

Non-Sequential Search, Screening Externalities and the Public Good Role of Recruitment Offices

Pieter A. Gautier

Free University Amsterdam

February 1999

Abstract:

This paper focuses on a congestion externality in the labor market which is caused by non sequential search. A model is presented in which unemployment is caused by selection delays of the employers. It is shown that the individual probability to get a job offer is increasing in the number of applications but that the aggregate hiring rate is decreasing in the average amount of applications per applicant. The congestion externality can be reduced by institutions which specialize in screening large groups of workers such as temporary employment agencies, labour offices, and head hunters. (JEL classification: C44, D62, J64, Keywords: Search Externalities, Queuing, Unemployment)

1 Introduction

In many labor market models the flow out of unemployment originates from the desire to search. A representative worker samples wage offers sequentially and decides on the basis of the sample obtained whether or not to continue searching, see e.g. Lucas and Prescott (1974). Alternatively, the matching models of Diamond (1992), Pissarides (1990) and others allow for search on two sides of the market. Those models have proved to be very useful and they have large explanatory power. There exists however a lot of micro-evidence that unemployment is more associated with "waiting" than with "searching". Moylan et al. (1982) report that most unemployed workers do not spend more than five hours a week on search activities while Van den Berg (1990) and others (summarized in Devine and Kiefer 1991) find that the fraction of job offers for which the offered wage exceeds the reservation wage is close to one¹.

This paper therefore takes a different route and presents a simple model in which unemployment is explicitly modelled as a state of waiting. In the model, unemployment is not caused by the fact that vacancies and workers cannot find each other but by the fact that it takes time for firms to process job applications and interviews. There are reasons to believe that those screening delays are at least as important as the contact delays which play a central role in the traditional matching models. Van Ours and Ridder (1992) show for example that the rate at which vacancies are filled up is low in the first few weeks and increases afterwards, while applicants mainly arrive in the first few weeks (in the last weeks the application arrival rate is even close to zero). This suggests that queuing and screening are important phenomena that deserve to be modelled seriously.

The distinction between search and wait unemployment is of course partly semantical but there are important differences between the approach in this paper and the ones followed in related flow models of

¹ The reason for this is that the job offer arrival rate is very low. When one rejects an offer, one has to wait for a long time to get another offer.

the labor market, like e.g. Pissarides (1990). First, the matching process is treated less as a blackbox since the effects of screening and selection on unemployment are explicitly taken into account. Second, I allow workers to search non-sequentially. Since employers can open as many vacancies as they want but workers can, in general, only occupy one job, it is only natural to allow workers to apply to more than one job. In the traditional search/matching models, where unemployment is caused by the fact that it takes time for vacancies and workers to find each other, the best strategy for a worker is to take the first acceptable job offer. In this framework that would not be a good strategy since when workers have to wait to be screened, it is in general optimal for them to apply for more than one job. The model here has in common with the search and matching models that it enables us to study the individual and the aggregate effects of (non sequential) job search in a unified framework. It will be shown for example that the individual hiring rate is increasing in the number of applications but that the aggregate hiring rate is decreasing in the average number of applications, although it is increasing in the number of job searchers.

The outline of the paper is as follows. Section 2 explains the main ideas of “wait unemployment” with a relatively simple model. To emphasize the differences with the traditional matching models, I assume that there are no contact delays at all. The equilibrium unemployment stock, the optimal screening intensity and the optimal amount of applications are derived both from a social welfare and from an individual point of view. It turns out that because of screening and congestion externalities, unemployment will be too high in equilibrium. Section 3 extends this model and allows for the possibility of rejecting applications and for free entry of screening points. In this case there can still be excessive congestion because the probability of a successful match only depends on the amount of job searchers and not on the amount of applications. It is true that more applications will increase the contact probability (and therefore stimulate entry of recruitment offices) but at the same time they will decrease the probability that the worker will accept a job offer by the same amount. Finally, it will be shown that institutions which centrally screen applicants can potentially reduce this externality.

2 A simple queuing model of the labor market

To get a clear idea about the differences between the non sequential search model in this paper and the traditional matching models, I will start with a simple version of the model which captures the main ideas of congestion unemployment. A key feature of the model is that before a worker can occupy a job he needs to be screened. Both worker and employer benefit from the screening process in the sense that it improves the match quality. A natural assumption is therefore that the productivity of a worker is an increasing function of screening time. Screening is costly however, the firm therefore faces a trade off between screening costs and match quality. The worker's problem is different. He has to decide how many applications he has to place, given the fact that applications are costly but also increase the probability to enter employment. An individual worker will however not take into account that each of his applications will increase the average waiting time of the other job searchers. Consequently, the number of applications will be too high in equilibrium. Thus, individual hazard rates are *increasing* in the number of applications while the aggregate hiring rate is *increasing* in the number of job searchers but it is *decreasing* in the number of applications per person. I will now formalize those ideas in a queuing framework.

Consider an economy with a large number of heterogeneous workers and one firm with a large number of (potential) job opportunities. In this economy, the workers need to be screened first before they can be matched with a vacancy. The arrival rate of job applicants (α) is assumed to be a Poisson stream and each applicant, i , can apply to v_i jobs. The screening rate is assumed to be exponentially distributed with parameter μ .

$$b(t) = \mu e^{-\mu t} \tag{1}$$

Where $b(t)$ is the probability density function (pdf) of screening time. If in a certain time interval, the arrival of applications (αv) exceeds the maximum number of applications which can be handled (μ), queues

of unemployed workers will be formed. This information is already sufficient to derive the steady state queue line distribution and the steady state level of unemployment in this economy, which is done below.

A feature of the steady state is that the inflow of applicants is equal to the outflow, see figure 1, which shows the simplifying case of $v=1$.

FIGURE 1 ABOUT HERE

Let P_n be the probability to find n persons in the system. Then in equilibrium, in a small time interval h , the probability of being in state n and leaving it, has to be equal to the probability of being in state $n-1$ or state $n+1$ and moving to state n . Hence, $P_n \alpha h + P_n \mu h = P_{n-1} \alpha h + P_{n+1} \mu h$ and $P_0 \alpha h = P_1 \mu h$. As $\sum P_n = 1$, there is a unique solution for P_n , which can be found by repeated substitution:

$$P_n = \left(\frac{\alpha}{\mu}\right)^n \left(1 - \frac{\alpha}{\mu}\right) \quad (2)$$

Given the fact that the average worker sends in v applications one can derive the following steady state level of unemployment which is equal to the expected amount of applications in the queue (L) divided by the average amount of applications per worker.²

$$U = \frac{L}{v} = \frac{\alpha}{\mu - v\alpha} \quad (3)$$

The necessary condition for an equilibrium to exist is that the average screening rate exceeds the average arrival rate of applications ($\mu > v\alpha$). In what follows, I assume that this condition is met. If this condition is not met, unemployment will go to infinity. The expected duration for an application to be processed (D) is given by $L/v\alpha$ or:

² See e.g. Gross and Harris (1985).

$$D = \frac{1}{(\mu - v\alpha)} \quad (4)$$

Note that both U and D are increasing in v and α . In particular when $v\alpha$ comes close to μ , unemployment will become very large. The hazard rate out of unemployment (h_i) for an individual worker who places v_i applications can be written as $(\alpha v_i/L)$, hence:

$$h_i = \frac{v_i}{v}(\mu - \alpha v) \quad (5)$$

Where v_i is the amount of applications of worker i . Thus, the individual hiring rate is increasing in v_i but the aggregate hiring rate is decreasing in v .

To determine the wage level, first assume that the productivity of a worker equals y/μ , which is increasing in the screening time $1/\mu$. This captures the idea that more screening effort of the firm leads to better matches. The flow cost of screening is given by k/μ^2 . If μ is large, little time per worker is spent on screening, as a result, screening costs and the output loss of foregone production will be small, but at the same time productivity will fall because workers and jobs will be matched relatively badly.

After the firm has picked a worker from the queue and has screened him to reveal his productivity, the wage bargaining starts. At that point, the firm only bargains with this one worker but the same solution can be supported by an aggregate bargaining between an employers organization and a union. For simplicity I assume that the match surplus is shared equally. Given the equilibrium wage, the individual workers and employers have to choose the optimal amount of applications and screening intensity.³ The equilibrium wage can then be derived in the standard way (first explored by Diamond (1982a)) by first giving an asset

³ I only implicitly take the heterogeneity of the labor force into account by assuming that there are benefits of screening. In other words, the workers do not differ in the potential output they can produce but their talents for different occupations do differ. Consequently, the firm needs to spend some time input to discover which job suits a particular worker best. Similar problems are also manifest in the ranking model of Blanchard and Diamond (1994) where there is no explicit reason why firms should rank workers in the first place. One can even argue that the search/matching models also suffer from this problem because it is hard to believe why search and matching frictions exist in the first place if all workers and jobs are identical.

value to each of the worker and job states. Those asset values will represent the threat points of the bargaining.⁴ The expected income stream for an unemployed worker can be denoted by:

$$rW_{U_i} = b - cv_i^2 + h_i[W_E - W_U] \quad (6)$$

Where cv_i^2 is the flow cost of the applications and h_i is given by (5)⁵. I will assume that the asset value of being employed equals:

$$rW_E = w - q[W_E - W_U] \quad (7)$$

Where q is the probability that the job is hit by an exogenous shock which destroys the job⁶. If the job losers would flow to unemployment, the additional assumption has to be made that in steady state equilibrium $qE=hU$. If they flow to non-participation, this assumption is not necessary. I will now turn to the employers side. The expected income stream of a filled job is assumed to be equal to:

$$rW_F = \frac{y}{\mu} - \frac{k}{\mu^2} - w - qW_F \quad (8)$$

According to the Nash solution of the bargaining, the wage is set in order to maximize the weighted product of the worker's and firm's surpluses which result from the job.⁷

$$(W_E - W_U)^\beta = W_F^{(1-\beta)} \quad (9)$$

where β denotes the worker's relative bargaining strength. To keep things as simple as possible, I will only

⁴ See also Hosios (1990) for an overview of the similarities between different matching models with non cooperative wage bargaining and an exposure on how other natural rate models of unemployment could be reformulated as matching-bargaining problems.

⁵ There are two points that deserve more attention. First, the idea behind the quadratic application cost is that it becomes increasingly more difficult to find a suitable vacancy. Second, a more natural assumption is that the application costs are made only once and are equal to $k\nu^2$. One can however transform those costs into a per period flow cost of $rk\nu^2$. Therefore, c can be interpreted as being equal to rk .

⁶ Here I assume that a job loser receives the utility of an unemployed worker. Alternatively one could assume that job losers will receive nothing or that jobs exist forever. Neither of those assumptions will change the main results.

⁷ For a justification of the Nash bargaining solution in this context see e.g. Binmore, Rubinstein and Wolinsky (1986).

consider the symmetric situation where β is equal to $\frac{1}{2}$. The first order condition implies then that:

$$W_E - W_U = W_F \quad (10)$$

Since the union maximizes the expected income stream of the *average* worker, the relevant threat point of the union becomes:

$$rW_U = b - cv^2 + h[W_E - W_U] \quad (11)$$

Where the hiring rate $h=(\mu-\alpha v)$. There are a number of ways to calculate the equilibrium wage, I will show one. From (11) and (7) an expression for $[W_E-W_U]$ can be obtained which can substituted together with (8) in the sharing rule (10). This results in:

$$w = \frac{(r + q)(b - cv^2) + (r+q+\mu-\alpha v)\left(\frac{y}{\mu} - \frac{k}{\mu^2}\right)}{2(r+q) + (\mu-\alpha v)} \quad (12)$$

Note that the wage is increasing in benefits (b) and screening costs (k), which improve the relative bargaining position of the workers. It is also increasing in y because a higher value of y implies a higher match surplus. Wages are decreasing in application costs (c) and the average amount of applications v (since $\mu - \alpha v > 0$). The effect of the screening time ($1/\mu$) on the wage is ambiguous. A longer screening time leads to more congestion and higher screening costs but it also improves the quality of the match and increases output.

The individual worker faces the problem to choose the amount of applications v_i which maximizes W_{U_i} , given the wage w . This optimal value v_i^* can be found as follows. First take the difference of (7) and (6) to get an expression for $[W_E-W_{U_i}]$, which can then be substituted in (6) again, and then maximize W_{U_i} with respect to v_i . The socially optimal number of applications per person v^* is determined in a different way. It can be obtained by maximizing (11) with respect to v . Below it will be shown that in equilibrium there

will be a tendency to "overapplicate".

Proposition: $v_i^* \geq v^*$ if $w > b$.

Proof. Workers will only apply for a job if the asset value of employment exceeds the asset value of unemployment, $[W_E - W_U] > 0$ ⁸. This implies that $dW_{U_i}/dh_i > 0$ (see (6)) and $dW_U/dh > 0$ (see (11)). In the market optimum, $dh/dv_i = (\mu - \alpha v)/v > 0$, see (5), while $dh/dv = -\alpha < 0$. From (11) and (12) it follows then that $dW_U/dv_i < 0$. The socially optimal expected income stream is decreasing in the total amount of applications, hence $v^* = 1$ (the minimum value of its domain). In the market optimum, maximizing (6) with respect to v_i gives,

$$v_i^* = \frac{-(r+q) + \sqrt{(r+q)^2 + \frac{(\mu - \alpha v)^2(w-b)}{cv}}}{\frac{(\mu - \alpha v)}{v}} \quad (13)$$

Note that $v_i^* > 0$ if $w > b$, and $v_i^* > 1$, if $w > c[1 + (2(r+q)v)/(\mu - \alpha v)] + b$.

It will be hard to design a contract which will internalize those costs because the determination of the wage and the worker's decision on the amount of applications are taken independently.

The firm's problem

The firm's problem is to determine the optimal screening rate μ^* , given the wage w . This optimal screening rate can be found by maximizing the value of a filled job with respect to μ . From (8) it follows then that $\mu^* = 2k/y$. From a worker's point of view, the optimal screening time is the value of μ which maximizes W_{U_i} . After evaluating the derivative of W_{U_i} with respect to μ at μ^* , it follows from (6) and (7) that:

⁸Otherwise we get the trivial solution $v_i = v^* = 0$.

$$\begin{aligned}
\frac{dW_{Ui}}{d\mu} \Big|_{\mu^*} &= \frac{\partial W_{Ui}}{\partial \mu} \Big|_{\mu^*} + \frac{\partial W_{Ui}}{\partial w} \frac{dw}{d\mu} \Big|_{\mu^*} \\
&= \frac{\frac{v_i(w-b+c)}{v}}{\left(r+q+\frac{v_i(\mu^*-v\alpha)}{v} \right)^2} + \frac{\frac{v_i(\mu^*-v\alpha)}{v}}{r+q+\frac{v_i(\mu^*-v\alpha)}{v}} \frac{dw}{d\mu} \Big|_{\mu^*}
\end{aligned} \tag{14}$$

The first term on the rhs of (14) is always positive while the sign of the second term depends on the sign of $dw/d\mu$. From (12), it follows however that $dw/d\mu|_{\mu^*} < 0$. Thus at the market equilibrium screening rate (μ^*), this second term at the rhs of (14) is negative. Hence, only if the effect of μ on w is relatively small, the employer's screening behavior causes a negative externality on the job searchers, because the waiting costs of the unemployed will not be internalized. Unemployment unambiguously increases however with screening time.

3 Extensions

In this section, the bare bones model of the previous section is extended by allowing firms to reject applications and by allowing screening points to enter freely. One can think of those screening points as being recruitment offices.⁹ The supply of those recruitment offices is basically determined by the same forces as the supply of vacancies in the matching model of Pissarides (1990) and hence both concepts are strongly related. The key difference is that a vacancy disappears as soon as it is occupied while recruitment offices in general remain open after a match. The easiest way to view those recruitment offices is that they all represent different firms with a large number of *potential* vacancies. The introduction of more than one screening point brings in a number of complications because the steady state queue line distribution depends on the ratio of workers and screening points now. It is assumed that every screening point has its

⁹ Under centralized screening, one can think about each screening point as representing an administrator who spends time and resources to interview the job applicants.

own job applications queue and that the employers will open up screening points till the marginal value of an additional screening point will be zero. In this extended model, the existence of a screening process is also easier to explain since one can now simply assume that a fraction of the applications does not fulfil the criteria for a certain occupation and that the firm needs time to separate the suitable from the unsuitable candidates.¹⁰ This extension does however give rise to another inefficiency because once a worker is found unsuitable for a particular occupation, this information is lost and the next firm has to spend screening time on this worker again. Central screening offices, head hunter agencies or temporary employment offices can play a public good role in the sense that they have to screen workers only once before they allocate them to suitable jobs. A second advantage of those institutions is that they can pool resources which leads to shorter worker queues. The role of those institutions is illustrated in the model by distinguishing between centralized and decentralized screening.

3.1 A queuing network of the labor market

Assume again that firms invite all applicants to enter the screening process and that the pdf of screening time is given by (1).¹¹ As in the previous section, intense screening is costly but leads to better matches, according to equation (8). When there are more screening points, the average waiting time of course decreases because the workers are distributed over more queues. Furthermore, assume that the firm determines the optimal screening rate in the same way as in the previous section.

The extended model makes a distinction between centralized and decentralized screening. The major differences between the two are summarized below. Under centralized screening, workers are distributed

¹⁰ Teyssiere (1996) finds for example that only 166 out of a sample of 2393 workers who got an interview (collected from a local employment office in Marseille) were hired after their first or second interview.

¹¹ This is a simplifying assumption which can easily be replaced by others. When firms invite for example a fixed number of applications, total screening time will be constant and "wait unemployment" will be less important. There will however still be "overapplication" in equilibrium.

equally over the screening points at each point in time while under decentralized screening this is not necessarily the case. Consequently, in the decentralized case, queues will only be formed when there are more applications than screening points. An interesting extension would be to endogenize the decision which queue to join. When there are different quality jobs that pay different wages (for example because jobs differ in the extent to which sunk investments can be appropriated ex post), then it is likely that the queues for high quality jobs will become longer (and the probability to get such a job will be lower). In equilibrium, the marginal benefits of waiting in a queue for a good job will be equal to the costs, in terms of a lower probability to get that job and a longer waiting period. When people have different preferences concerning the disutility of waiting because there are for example differences in initial wealth, this may give rise to mismatches. High skilled workers with little initial wealth may in that case be forced to join the short queues associated with low quality jobs (since they cannot afford to remain unemployed for a long period)¹². For the moment I will leave this for what it is.

In the decentralized case, it can be decided that the worker is unsuitable for the occupation he applied for. Let p be the probability of getting an acceptable job offer and $(1-p)$ the probability to be rejected. Then, after the unemployed worker has decided to how many jobs he wants to apply, he faces a probability $1-(1-p)^n$ that he is invited into the screening process and a probability $(1-p)^n$ that he is rejected. If he is rejected, he will flow back to the pool of unemployed.

3.2 Steady State properties of a centralized and a decentralized system

This section shows how the steady state properties differ between a centralized and a decentralized system. In what follows I will use subscript c to denote the centralized case and subscript d to denote the decentralized case. For simplicity it is assumed that the acceptance rate in the decentralized case=1. The

¹² This will not happen under perfect capital markets with complete costless information.

results will not change qualitatively when $0 < p^d < p^c < 1$. In the centralized case (one pooled system) the arrival rate is given by:

$$\alpha^c = v(\tau_1 + \tau_2) \quad (15)$$

where $v\tau_1$ and $v\tau_2$ are the applications of non participants and fired workers respectively. In the decentralized case the arrival rate is given by:

$$\alpha^d = \frac{v(\tau_1 + \tau_2)}{pS} \quad (16)$$

An analogue to this system is the difference between n different shops with 1 cash deck, an arrival rate of α/n for each shop and a probability p that one's credit card is not accepted in a particular shop which forces the customer to join the queue of another shop, and a system with only one shop with n cash decks, an arrival rate α and where all credit cards are accepted. Even if p is 1, queues will be shorter in the second case because customers will distribute themselves evenly over the queues. The steady state queue line distribution and unemployment stock for the decentralized case can be calculated in a straightforward way analogue to (2) and (3). This results in:

$$U^d = \frac{v(\tau_1 + \tau_2)}{pS\mu - v(\tau_1 + \tau_2)} \quad (17)$$

where S is the amount of screening points in equilibrium, which will be derived in section 3.3. Under centralized screening, workers only need to apply once.¹³

For the decentralized case, things are a little more complicated. From (15), (2) and appendix 1, which shows how can be dealt with "multiple screening points", the joint steady state queue line distribution can be derived. The system is totally described by the waiting times, the number of screening points (S), the

¹³ This is again an extreme case. A worker can of course choose to register at different temporary employment agencies. What matters here is that this causes less congestion than when he would apply directly to different jobs. We can therefore also interpret v as the difference in applications under centralized and decentralized screening.

number of applications, and the inflow into unemployment.

$$P(k) = \left(\frac{\left(\frac{(\tau_1 + \tau_2)}{\mu} \right)^k}{S^{k-S} S!} \right) \left(\sum_{k=0}^{S-1} \frac{\left(\frac{(\tau_1 + \tau_2)}{\mu} \right)^k}{k!} + \frac{\left(\frac{(\tau_1 + \tau_2)}{\mu} \right)^S}{S! \left(1 - \frac{(\tau_1 + \tau_2)}{S\mu} \right)} \right)^{-1} \quad (18)$$

From (18), the steady state unemployment rate can be derived (appendix 1 shows how).

$$U^c = \frac{\left(\frac{(\tau_1 + \tau_2)}{\mu} \right)^{S+1}}{\left(S - \frac{\tau_1 + \tau_2}{\mu} \right)^2 (S-1)!} \left(\sum_{n=0}^{S-1} \frac{\left(\frac{(\tau_1 + \tau_2)}{\mu} \right)^n}{n!} + \frac{\left(\frac{(\tau_1 + \tau_2)}{\mu} \right)^S}{S! \left(1 - \frac{\tau_1 + \tau_2}{S\mu} \right)} \right)^{-1} + \frac{\tau_1 + \tau_2}{\mu} \quad (19)$$

This is the sum of applications "waiting" at the different screening points. Steady state unemployment increases with the number of job searchers, the average amount of applications, gross in and outflow, screening time and, although it is not obvious from this equation, it decreases with the number of screening points (S).

The average individual hiring rate is given by the amount of applications times the outflow rate (in persons) divided by the total amount of applicants (νU^d in the decentralized case and U^c in the centralized case).

$$h_i^d = \frac{\nu_i \tau_1 + \tau_2}{\nu U^d(\nu)}, \quad h_i^c = \frac{\tau_1 + \tau_2}{U^c} \quad (20)$$

Note again that the individual hiring rate in the decentralized screening regime is increasing in ν while the aggregate hiring rate, $(\tau_1 + \tau_2)/(U^d(\nu))$ is decreasing in ν because ν increases steady state unemployment.

In the centralized screening regime, workers have no incentives to choose $\nu > 1$, hence the aggregate hiring

rate is larger in that case. In the next section, some simulations are shown for both regimes. The simulations are carried out in a way that a distinction can be made between the public good role and the pooling role of centralized recruitment offices.

3.3 Simulations

To get some feeling about the working of the model, a small numerical example is given in this section. In reality, search strategies and escape probabilities differ between sectors and types of labor. For low skilled workers, the search and application time will probably be very short while for high skilled labor, search and application time will be longer. This model ignores all those issues. Nevertheless, it is instructive to see how large steady state unemployment becomes when the arrival rate of applications approaches the screening rate.

In this example, τ_1 and τ_2 are set at values equal to their average (x1000) in the Netherlands¹⁴, see Gautier and Broersma (1994), p , the probability of a match after a contact has been made is set at 0.15, this value is based on estimates of Lindeboom et al. (1993). Finally, μ is chosen in a way that employers can just handle an average of 3 applications per worker per month. I distinguish between two cases, one in which screening takes place completely decentralized and in the other case I consider one fully pooled institution which screens all applicants and assigns them to different jobs. The public good role is modelled by reducing the rejection probability p , capturing the idea that if a worker is found unsuitable for occupation i , he might still be the right candidate for occupation j . The other advantage of a centralized screening office arises from the fact that there is a particular form of increasing returns to scale, due to the pooling of resources. This leads to a better allocation of workers over the different queues. In the most efficient case, when every worker is screened centrally and no workers are rejected, v is 1 and p is 1. Table 1 shows

¹⁴ τ_2 is set at 287 (the average yearly flow from employment to unemployment (70-91)) and τ_1 is set at 287 (the average unemployment inflow minus the flow from employment to unemployment).

that steady state unemployment is in this case negligible. On the other hand in the decentralized case, when there are 50000 decentralized screening points, all workers apply to 3 jobs and face a rejection probability of 85%, unemployment can get as high as 3 mln.

Table 1 Expected unemployment (x1000) for different values of ν and p .

ν	p	U^c	U^d
1	0.15	3.3	3014.0
2	0.50	1.0	4.3
3	1.00	0.5	1.8
1	0.15	3.3	19.8
2	0.50	1.0	2.5
3	1.00	0.5	1.1
1	0.15	3.3	5.0
2	0.50	1.0	1.1
3	1.00	0.5	0.5

$\mu_{cen}=1057.5, \mu_{decen}=211.5, S=50$ (x1000).

To get an idea of the separate effects of pooling and the "public goods" role of recruitment offices, I introduced exactly the same rejection probability for the decentralized case. Consider for example the case where ν is equal to two. When the acceptance probability, p , is 0.15, the advantages of pooling are much larger than when p is 0.5. In other words, the closer the arrival rate of applicants comes to the screening capacity, the larger the advantages of central recruitment agencies are.

3.4 The equilibrium wage

In the previous sections, the amount of screening points, S , was still treated exogenously. This section derives the amount of screening points under free entry in the decentralized case. This can however not be done properly without taking wages into account. Wages are assumed to be determined in the same way as in section 2.

To determine the equilibrium wage, one can basically use equations (6),(7), (8) and (10) again. Where h^d_i is now given by (20), and $q = \tau_2/E$. Thus in equilibrium,

$$W_F = \frac{\frac{y}{\mu} - \frac{k}{\mu^2} - w^d}{r+q} = (W_E - W_U) = \frac{w^d - b + cv^2}{r+q + \frac{\tau_1 + \tau_2}{U}} \quad (21)$$

Solving for the wage w yields¹⁵:

$$w^d = \frac{(r+q)(b - cv^2) + \left(r+q + \frac{\tau_1 + \tau_2}{U^d}\right) \left(\frac{y}{\mu} - \frac{k}{\mu^2}\right)}{2(r+q) + \frac{\tau_1 + \tau_2}{U^d}} \quad (22)$$

Note that the wage depends positively on production and unemployment and negatively on the application and screening costs. Under centralized screening, a higher wage would result because the outside option of workers improves since $U^c < U^d$ and less application costs have to be made. Thus, part of the efficiency gains of centralized recruitment will be appropriated by the workers in the form of higher wages.

3.5 The amount of screening points in equilibrium

To derive the amount of recruitment offices in decentralized equilibrium, I will assume in addition that the flow costs of opening a recruitment office are equal to c_S . For an individual recruitment office i , the fraction of acceptable applications will be equal to the total number of applications which arrive at recruitment office i , divided by the total number of recruitment offices. Since only a fraction $1/\nu$ of the applicants will accept firm i 's job offer, this number has to be divided by ν .

$$rW_S = -c_S + \frac{p\nu(\tau_1 + \tau_2)}{\nu S} W_F = 0 \quad (23)$$

I have assumed here that in an efficient equilibrium all profit opportunities will be used, hence $rW_S = 0$. Solving for S gives:

¹⁵ The costs associated with the screening points are sunk and play no role in the wage bargaining.

$$S = \frac{(\tau_1 + \tau_2)p \left(\frac{y}{\mu} - \frac{k}{\mu^2} - w^d \right)}{c_v(r+q)} \quad (24)$$

Note that the supply of screening points increases with the number of applicants. The effect of the average number of applications per applicant on the supply of screening points is more complicated however. There is no direct effect because jobs are filled by persons and not by applications. There is however a second order effect which goes through the wage. If the average application rate is large, the state of unemployment becomes less attractive and the worker or union will take this into account by accepting a lower wage. This lower wage will result in a higher match surplus for the employer and consequently in a higher equilibrium stock of screening points.

$$\frac{dS}{dv} = \frac{\partial S}{\partial w} \frac{dw}{dv} = \frac{p_2 c_v (\tau_1 + \tau_2)}{c_s \left((r+q) + \frac{(\tau_1 + \tau_2)}{U^d} \right)} > 0 \quad (25)$$

The social benefits of an extra screening point are in terms of less congestion and unemployment because it will speed up the rate at which matches take place. This follows directly from (17). Individual firms will however not internalize those positive external effects of more screening points. Also note from (24) that if (pooled) screening points would be assigned according to the same free entry rule, more screening points would be opened because the expected benefits would be higher. The main reason for this is that screening always leads to a match in the centralized regime. The total effect will be positive even though the fact that $w^f > w^d$ dampens this effect.¹⁶

4 Final Remarks

This paper focuses on a particular type of congestion externality namely one in which an increase of the search intensity of one person, increases the average waiting time of the others and hence reduces the expected benefits of search. Simulations show that small changes in the screening and application intensity

¹⁶ Recall that it was assumed that employers and workers split the match surplus equally. This means that the efficiency gains partly result in higher returns of a filled job and partly into a higher wage. Under free entry, the higher asset value of a filled job will induce additional entry.

at the micro level can result in large unemployment at the aggregate level. The labor market was portrayed as a queuing model in which unemployment duration is for a large part out of control of the worker and in the hands of other searchers and employers. Coordination failures with respect to the search intensity of workers lead to an equilibrium where workers who decide to search, apply for too many jobs and because of that, the rewards to search are lower than the socially optimal rewards. Institutions which specialize in screening like head hunters and temporary employment offices fulfil a useful role in reducing unemployment because they are able to pool resources and they have a public good role in the sense that the information of a screened applicant is not lost after he is rejected to a particular job.

In reality there may of course be good other reasons (i.e. increased competition) for decentralized screening so that the "true" optimum will be somewhere in between totally decentralized and totally centralized screening. The results of this paper are also consistent with the observation that large firms pay higher wages than small firms. Large firms typically have large recruitment offices and many different occupations so that the probability that screening a given worker will lead to an acceptable match is higher for those firms. It is likely that part of those efficiency gains will be appropriated by the worker in the form of a higher wage.

References

- Abbring J.H., J.C. van Ours** (1994), Sequential or non-sequential employers' search?, *Economics Letters* 44, pp. 323-328.
- Berg van den G.J.** (1990), Search Behaviour, Transitions to Nonparticipation and the Duration of Unemployment", *Economic Journal*, 100, pp. 842-65.
- Binmore K., A. Rubinstein and A. Wolinsky** (1986), The Nash Bargaining Solution in Economic Modelling, *Rand Journal of Economics*, 17, pp. 176-88.
- Blanchard O.J. and P. Diamond** (1994), Ranking, Unemployment Duration, and Wages, *Review of Economic Studies*, 61, pp. 417-434.
- Burke, P.J.** (1956), The Output of a Queuing System, *Operations Research*, 4, pp. 699-704.
- Devine T.J. and N. Kiefer** (1991), *Empirical Labor Economics*, Oxford University Press.
- Diamond P.A.** (1982) Wage determination and efficiency in search equilibrium, *Review of Economic Studies*, 49, pp. 217-227.
- Gautier P.A. and L. Broersma** (1994), The Timing of Labor Reallocation and the Business Cycle, Mimeo, Tinbergen Institute.
- Gross D. and C.M. Harris** (1985), *Fundamentals of Queuing Theory*, Wiley and Sons, NY.
- Hosios, A.J.** (1990), On the efficiency of Matching and Related Models of Search and Unemployment, *Review of Economic Studies*, 57, pp. 279-298.
- Jackson J.R.** (1957), Networks of Waiting Lines, *Operations Research*, Vol. 5, pp.518-521.
- Layard R.S. Nickell, R. Jackman.** (1991), "Unemployment", Oxford University Press.
- Lemoine A.J.**, (1977), Networks of Queues- A survey of Equilibrium Analysis, *Management Science*, 24, pp. 464-481.
- Lindeboom M., J.C. van Ours and G. Renes**, (1993), Matching Job Seekers and Vacancies, in: *Labor Demand and Equilibrium Wage Formation*, Eds. J.C. van Ours, G.A. Pfann, and G. Ridder, Amsterdam,

North Holland, pp. 279-296.

Lucas R.E. and E.C. Prescott (1974), Equilibrium Search and Unemployment, *Journal of Economic Theory*, pp. 188-209.

Moylan S., Millar, J, and Davies, R. (1982), "Unemployment-The Year After", Dep. of Employment, *Employment Gazette*, 90: pp.334-40.

Ours J.C. van and G. Ridder (1992), "Vacancies and the Recruitment of new Employees, *Journal of Labor Economics* 10: 138-155

Pissarides C.A., (1990), *Equilibrium Unemployment Theory*, (Basil Blackwell, London).

Teyssière, (1996), Matching Processes in the Labour Market. An Econometric Study, *Labour Economics* 2, pp.421-435.

Appendix

Multiple screening points

This appendix describes the steady state properties of a centralized recruitment office with multiple screening points. It will be assumed that all screening points are identical and that each application joins the smallest queue. To derive the steady state properties of a multiple screening system the balance equation method will be used again. In a very small time interval h , the following relations hold in the steady state:

$$P_0 \alpha h = P_1 \mu h$$

$$P_1 \alpha h + P_1 \mu h = P_0 \alpha h + 2P_2 \mu h$$

$$P_2 \alpha h + 2P_2 \mu h = P_1 \alpha h + 3P_3 \mu h$$

· · · · ·

· · · · ·

· · · · ·

$$P_n \alpha h + n P_n \mu h = P_{n-1} \alpha h + (n+1) P_{n+1} \mu h \quad 0 < n \leq S - 1$$

$$P_n \alpha h + S P_n \mu h = P_{n-1} \alpha h + S P_{n+1} \mu h \quad n \geq S$$

Consider the following two cases. (1) There are less applications than screening points and no queues will be formed. (2) There are more applications than screening points. Case 2 is the most relevant case. In the sequel it is assumed that $0 < \alpha/\beta < 1$.

Case 1. $0 < n \leq S-1$. The steady state queue line distribution is given by:

$$P_n = \frac{\left(\frac{\alpha}{\mu}\right)^n}{n!} P_0$$

Case 2. $n \geq S$ The steady state queue line distribution is given by:

$$P_n = \frac{\left(\frac{\alpha}{\mu}\right)^n}{S^{n-S}S!}P_0 = \frac{\left(\frac{\alpha}{\mu}\right)^S \left(\frac{\alpha}{\mu}\right)^{n-S}}{S! S^{n-S}}P_0$$

One can write P_0 in terms of α , μ and S by using the fact that all chances have to sum up to 1. This gives

$$P_0 \left(1 + \sum_{n=1}^{S-1} \frac{\left(\frac{\alpha}{\mu}\right)^n}{n!} + \frac{\left(\frac{\alpha}{\mu}\right)^S}{S!} \sum_{n=S}^{\infty} \left(\frac{\alpha}{S\mu}\right)^{n-S} \right) = 1$$

Hence,

$$P_0 = \left(\sum_{n=0}^{S-1} \frac{\left(\frac{\alpha}{\mu}\right)^n}{n!} + \frac{\left(\frac{\alpha}{\mu}\right)^S}{S! \left(1 - \frac{\alpha}{S\mu}\right)} \right)^{-1}$$

$$\text{Since } \sum_{n=S}^{\infty} \left(\frac{\alpha}{S\mu}\right)^{n-S} = \left(\frac{1}{1 - \frac{\alpha}{S\mu}} \right)$$

In the steady state, the expected number of units in the system (L) is equal to the sum of the average number of waiters (L_q) and the average number of people that is being screened (α/μ).

$$L = \frac{\frac{\alpha}{S\mu}}{1 - \frac{\alpha}{S\mu}} P(n \geq S) + \frac{\alpha}{\mu} = \frac{\left(\frac{\alpha}{\mu}\right)^{S+1}}{\left(S - \frac{\alpha}{\mu}\right)^2 (S-1)!} P_0 + \frac{\alpha}{\mu}$$

Figures

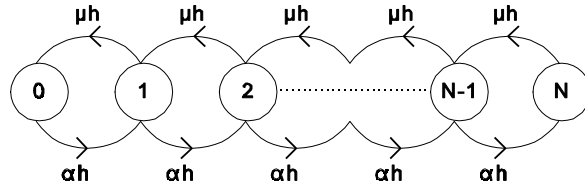


Figure 1 Transition diagram