Queueing and Searching

Michael Sattinger*
Department of Economics
University at Albany
Albany, NY 12222, USA
Email m.sattinger@albany.edu
Phone: (518) 442-4761
Keywords: Queueing, Search, Unemployment

October 18, 2007

Abstract

This paper applies queueing theory to derive the equilibrium of a labor market with frictions. Queueing arises when workers can get jobs by waiting at a firm for a position to open. Queueing theory provides expressions for the expected numbers of vacancies, searching workers and queueing workers. As the ratio of workers to firms increases, unemployed workers shift from searching for vacancies to waiting in queues.

1 Introduction

Queueing arises when individuals are not able immediately to obtain services or goods they seek and instead must wait while others are served or receive the goods before them. Queueing arises naturally in circumstances where sellers' capacity to provide services is limited or when sellers cannot keep a sufficient inventory of goods. Queueing theory can be used to describe

*The author is indebted to John Kennan, Adrian Masters and Robert Shimer for helpful comments. Any errors are the responsibility of the author.

the outcomes of processes using simple assumptions of agent behavior or it can be embedded in market contexts where agent behavior is endogenous. With different perspectives, statistics and probability, operations research and economics all contribute to queueing theory.

This paper considers the possible application of queueing theory in labor contexts and relates the subject to searching and matching. Basic questions include whether queueing phenomena occur in labor markets in any form and whether institutions that promote queueing could increase the efficiency of labor markets. Further questions include whether queueing theory can provide a microeconomic foundation for aggregate labor market phenomena and whether it can contribute to matching models. In addition to waiting unemployed in a queue for a job to open up at a firm, queueing in a labor market could take the form of queueing within a firm, waiting for a job in an industry, and waiting or queueing while employed at another firm.

Queueing theory as a discipline was developed by Agner Krarup Erlang.[1] Erlang began his career as a schoolteacher but continued his studies of mathematics and probability. Dr. F. Johannsen wrote two early papers providing approximate solutions to outcomes in telephone exchanges ("Waiting Times and Number of Calls," in 1907, and "Busy," in 1908). In 1908, Johannsen asked Erlang to treat the problems mathematically. The Copenhagen Telephone Company appointed Erlang as head of a research laboratory on telephone problems. In 1909, Erlang demonstrated that telephone calls arrive according to a Poisson process ("The Theory of Probabilities and Telephone Conversations"). In 1917, he wrote his most significant paper on the subject, "Solution of Some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges."

As a background to extensions of queueing theory, consider the single server problem that is often presented in texts on probability and statistics. Buyers arrive at a firm and wait in line (queue) until they are served, obeying the queue discipline "First In First Out."[2] With a single server, buyers do

[1] E. Brockmeyer, H.L. Halstrom and Arne Jensen (1948) have written a book about the life and works of A.K. Erlang, available online at oldwww.com.dtu/teletraffic/Erlang.html.

[2] See Cox and Smith, 1961; Gnedenko and Kovalenko, 1989; and Gross and Harris, 1985, for standard treatments of queueing theory. Basic treatment of queueing theory can also be found in texts on probability and stochastic processes such as Feller, 1957; Karlin, 1966; Karlin and Taylor, 1981; Saaty, 1961; and Taylor and Karlin, 1984. Hassin and Haviv, 2003, present recent developments in queueing theory including applications of game theory.

not search for different queues or "balk" by refusing to enter the queue when it is too long. Suppose the arrival of buyers is a Poisson process with average rate of arrival α. Suppose the rate at which buyers get the good when they get to the head of the queue is σ. The traffic intensity of the process is $\rho = \alpha/\sigma$ (defined in honor of Erlang after his death in 1929). It can be shown that the portion of time the server is idle (with no one in the queue) is $1 - \rho$, and the expected number of buyers waiting in the queue is $\rho/(1-\rho)$. Thus the queueing process generates a relation between the number of buyers in the queue, the proportion of time the server is idle, and sales (given by $\rho\sigma = \alpha$). This relation is equivalent to the "matching function" in labor economics, which relates the number of unemployed workers, the number of vacancies and the number of matches. This relation is inherent in the queueing process and does not rely on any assumptions about searching behavior of buyers. (Margaret Stevens, 2007, goes significantly beyond this by providing a queueing theoretic foundation for a specific labor market matching function based on behavioral assumptions.) The queueing process also generates frictions in the form of positive probabilities of idle servers and queueing buyers. (With multiple servers and no balking by changing queues, there will be simultaneous server idle time and queueing buyers.)

Now consider how going beyond the single server model begins to involve economic behavioral assumptions. Extending the single server model to multiple servers introduces the problem of how buyers find queues. Buyers must then find a queue to enter, and may engage in search among servers. With alternative servers, buyers may choose not to stay in a queue (i.e., they may "balk"). Since servers suffer losses from idle time, they can offer a lower price for the service or good in order to attract arrivals (consequences of a lower price depend on the information available to buyers). This introduces a pricing mechanism into the queueing process. Other behavior that is relevant to the queueing process includes action to change position in the queue depending on the queue discipline; sellers' decisions to enter the market with an additional server; and sellers' decisions to offer more attractive combinations of queue length and price. Game theory enters in priority auctions, in which a person in the queue bids to advance in the queue (Moldovanu and Kittsteiner, 2005). There is therefore a rich interplay between the economic context, assumptions about the queueing process, and economic behavior. In this way, a market equilibrium, with prices and quantities, can be determined for a market with queueing.[3]

Additionally, queueing theory can contribute to the analysis of market disequilibrium. In a Walrasian market model with an auctioneer, no trades occur out of equilibrium. If prices are such that markets do not clear, the auctioneer adjusts the prices until the equilibrium allocation arises, at which point trade commences. With price determination linked to allocation, the Walrasian model is uninformative about disequilibrium behavior or how markets actually clear (Dennis Carlton, 1989; Sattinger, 2002b). In contrast, a market model with queueing operates to allocate goods or services whether or not prices are in equilibrium. Decentralized choice of queues by buyers generates compensating expected wait times. As a consequence, a market with queueing involves one mechanism that allocates goods or services in all circumstances and a separate mechanism to change prices over time. The significance of this separation of pricing from the allocation mechanism is that a market with queueing can be used to study the consequences of price dispersion and the dynamics of price adjustment over time.

Queueing can be applied in a natural manner to labor markets since labor is a non-storable service. Either firms could queue at workers (the providers of services) or workers could queue at firms. It is possible to construct a queueing model of a labor market completely without search, depending on assumptions. For example, with standard assumptions for the rest of the model, suppose workers have full information about queue lengths at firms. Unless queues increase in length indefinitely, there must always be vacancies, so workers would never choose a firm with a worker in queue. If the number of workers is sufficiently high that some queueing will occur, both vacancies and unemployment (in the form of worker queueing) could occur simultaneously without any search taking place.[4]

However, since search occurs naturally in a multi-server queueing model without full information, it is more appropriate to construct a model with both queueing and searching. The following schematic shows one way in which queueing could occur in a labor market. Workers leave firms at a fixed rate. Firms can allow workers to enter a queue to take the next available job at the firm. If there are workers in the queue, a firm can avoid vacancies.

[3] See Allon and Federgruen, 2004; Knudsen, 1972; Leeman, 1964; Naor, 1969; Sattinger, 2002a; and Stahl, 1987, for applications of queueing in market contexts.

[4] The model developed by Lagos (2000) in the context of taxicabs and locations generates frictions endogenously. The model, which exhibits some aspects of common resource problems, would be well suited to a queueing theoretic treatment.

Otherwise, with no workers in the queue, a firm would experience vacancies when workers leave. Unemployed workers search among firms either for a vacancy or for a queue that is not too long. With fairly straightforward assumptions, it is possible to obtain analytical expressions for the expected numbers of firm vacancies, searching workers and queueing workers for a given aggregate ratio of workers to firms. Queueing modifies the results of a labor market with only search behavior. At low ratios of workers to firms, the labor market operates in the same manner whether or not queueing occurs, since vacancies would be common and queues would be rare. At high ratios of workers per firm (and with high unemployment), most unemployed workers would be waiting in queues, and vacancies would be substantially reduced by the presence of queues. Most workers would be searching for queues to join rather than vacancies.

In addition to modifying the outcomes in a search and matching model, queueing methodology can be used to generalize the standard model without requiring queueing. In the standard model, a job is coextensive with a firm. This greatly simplifies the analysis and is justified by assuming that every job is different. As a consequence of the one worker-one job assumption, jobs are in one of two states, filled or vacant, and the proportions of time in each are easily expressed in terms of the matching function and ratios of workers to jobs. However, it is unnecessary for the standard model. A firm with n workers will have $n + 1$ states depending on the number of vacancies. The proportions of time in each of these states can be calculated using methods of queueing theory. (If there are also queues, the number of states will be greater.) Then it is possible to dispense with the fixed proportions technology of one worker per firm.

As an alternative to the development in this paper, it is possible to consider queueing as the result of on-the-job search. Currently employed workers could queue at firms without vacancies, prepared to shift to the firm if a vacancy arises. In that way, a firm could avoid the cost of a vacancy, and workers could wait in a queue for a higher-paying job without losing their current income.

The next section develops the queueing theory of a symmetric equilibrium model of labor market equilibrium. This includes determination of the proportion of time firms have vacancies and queues, the maximum queue length, the proportion of worker times spent employed, searching and queueing, the firm arrival rate given the ratio of workers to firms, and the wage rate. Section 3 shows how a labor market with queueing generates a dual wage equilibrium

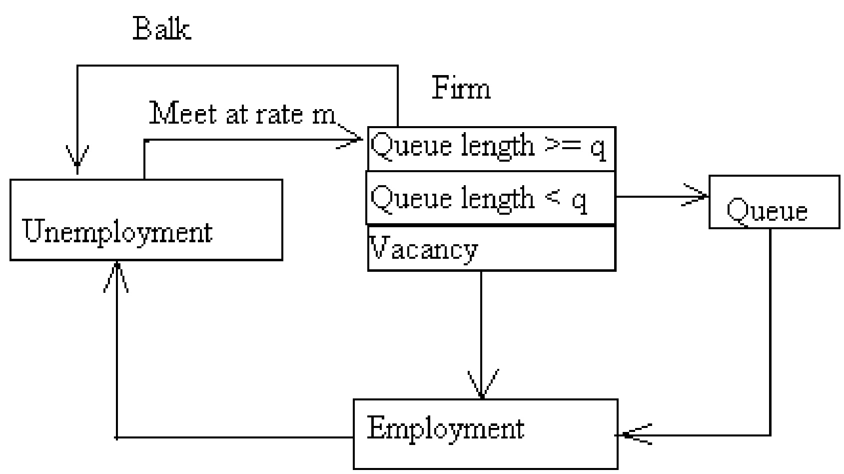

Figure 1: Schematic of Labor Market with Queueing

at higher arrival rates.[5] Section 4 describes the operation of a labor market with queueing. In particular, the operation changes qualitatively as the ratio of workers to firms increases, shifting from searching to queueing. Section 5 summarizes conclusions, discusses applications, and describes how queueing workers would appear in the Current Population Survey.

2 Symmetric Equilibrium Model

2.1 Overview and Definitions

This section describes the steady-state symmetric equilibrium in a model of labor queues in continuous time. The next section will extend these results to equilibria that are not symmetric, specifically with two arrival rates at firms, two wage rates and two unemployment rates. Nonsymmetric equilibria arise at higher ratios of workers per firm.

Production at a firm is proportional to employment up to a maximum n, the same for all firms. A firm employing j workers has $n-j$ vacancies. Unemployed workers may contact the firm to see if there is a vacancy and can also decide to enter a queue to wait for a job if the queue is not too long. Suppose workers in a queue are hired in the order that they enter the queue (i.e., the queue discipline is First In First Out). Define the balk level q as the maximum length of a queue, so that an unemployed worker would refuse to enter a queue of length q or longer. Let α be the arrival rate of unemployed workers at a firm looking for vacancies or a queue shorter than q. Suppose there are N_W workers and N_F firms. The arrival rate will be determined endogenously from the ratio of workers to firms in Section 2.5. Let γ be the rate at which employed workers depart jobs. Assume γ is the same for all workers and firms. Let $-i$ be the firm's state when there are i vacancies, $i = 1, n$, and let p_{-i} be the (long run) proportion of time a firm is in this state. Let i be the firm's state when it has i unemployed workers in the queue for a job, $i = 1, q$, and let p_i be the proportion of time the firm is in this state. Let 0 be the state when the firm has n workers with no workers in the queue and let p_0 be the proportion of time a firm is in this state. The proportions p_i, $i = -n, ...q$, are determined by the arrival and departure rates α and γ and are derived in Section 2.2. At higher arrival rates α, firms are

[5] Gaumont, Schindler and Wright, 2006, review labor markets that generate dual wage equilibria.

less likely to have vacancies and more unemployed workers are in queues.

Workers maximize their present discounted asset values from future employment with an infinite horizon. They know wage rates and expected queue lengths at different firms (which would be identical in a symmetric equilibrium) and can calculate the value of searching at different firms. Workers do not know actual queue lengths and therefore do not direct their search towards firms with shorter actual queues. Firms offer wages that do not vary with queue length.

Among firms paying the same wage rate, workers optimally choose the balk level q by comparing the expected time until they get a job if they continue searching with the expected time if they join a queue with $q-1$ workers in queue ahead of them. Let V_E, V_S and V_Q be the expected asset values for a worker entering the states of employment, searching and queueing, respectively. These asset values are derived in Section 2.6. An individual firm takes the asset value of searching, V_S, as given. An individual firm can choose the wage rate it pays but then the arrival rate at the firm will adjust to a level such that the value of searching at the firm is the same as searching in the rest of the labor market, given by V_S, with balk level q adjusting. The firm optimally chooses the wage subject to the constraint on the arrival rate. It is then possible to determine the symmetric equilibrium of a labor market consisting of identical workers and firms.

2.2 Steady-State Firm Proportions of Time in Each State

Following standard procedures in queueing theory, proportions of time a firm is in each state can be found from transition rates into and out of particular states. Beginning with the state of no workers (state $-n$), the average transition rates for a firm into and out of the state must be equal:

$$\gamma p_{1-n} = \alpha p_{-n} \tag{1}$$

The average transition rate of firms into state $-n$ equals the departure rate γ times the proportion of time the firm is in state $1-n$. The average transition rate of workers out of state $-n$ equals the arrival rate of workers α times the proportion of time the firm is in state $-n$.

The average transition rate into state $1-n$ arises from worker departures when the firm is in state $2-n$ and arrivals when the firm is in state $-n$.

The average transition rate out of state $1-n$ arises from worker arrivals or departures. Setting the flows equal yields

$$2\gamma p_{2-n} + \alpha p_{-n} = \alpha p_{1-n} + \gamma p_{1-n} \tag{2}$$

Similarly, the average transition rates into and out of state $j-n$, with $n-j$ vacancies, are

$$(j+1)\gamma p_{j+1-n} + \alpha p_{j-1-n} = \alpha p_{j-n} + j\gamma p_{j-n} \tag{3}$$

The firm moves into state $j-n$ from state $j+1-n$ when any of the $j+1$ workers depart, or from state $j-1-n$ when a worker arrives. The firm moves out of state $j-n$ when a worker arrives or when any of the j workers departs. In state -1, with one vacancy, the average transition rates are

$$n\gamma p_0 + \alpha p_{-2} = \alpha p_{-1} + (n-1)\gamma p_{-1} \tag{4}$$

When the firm has workers in the queue, n workers are employed. Then the average transition rates in states 0 and 1 are

$$n\gamma p_1 + \alpha p_{-1} = \alpha p_0 + n\gamma p_0 \tag{5}$$

$$n\gamma p_2 + \alpha p_0 = \alpha p_1 + n\gamma p_1 \tag{6}$$

In state $j > 0$,

$$n\gamma p_{j+1} + \alpha p_{j-1} = \alpha p_j + n\gamma p_j \tag{7}$$

and in state q

$$\alpha p_{q-1} = n\gamma p_q \tag{8}$$

The equations 1 through 8 can be solved for the proportions p_i, $i = -n, q$. First, the proportions of time that the firm has vacancies can be expressed in terms of p_{-n}. From 1,

$$p_{1-n} = \frac{\alpha}{\gamma} p_{-n} \tag{9}$$

Combining 1 with 2 yields

$$2\gamma p_{2-n} + \alpha p_{-n} + \gamma p_{1-n} = \alpha p_{1-n} + \gamma p_{1-n} + \alpha p_{-n} \tag{10}$$

Canceling out terms yields

$$p_{2-n} = \frac{1}{2}\frac{\alpha}{\gamma}p_{1-n} = \frac{1}{2}\left(\frac{\alpha}{\gamma}\right)^2 p_{-n} \tag{11}$$

Similarly, combining equations for successive numbers of vacancies yields

$$p_{j+1-n} = \frac{1}{j+1}\frac{\alpha}{\gamma}p_{j-n} = \frac{1}{(j+1)!}\left(\frac{\alpha}{\gamma}\right)^{j+1} p_{-n} \tag{12}$$

From 4

$$p_0 = \frac{1}{n}\frac{\alpha}{\gamma}p_{-1} = \frac{1}{n!}\left(\frac{\alpha}{\gamma}\right)^n p_{-n} \tag{13}$$

In the second step, the proportions of time with workers in the queue can be expressed in terms of p_q. Beginning with 8,

$$p_{q-1} = n\frac{\gamma}{\alpha}p_q \tag{14}$$

Then at successively lower numbers of workers in the queue,

$$p_j = n\frac{\gamma}{\alpha}p_{j+1} = n^{q-j}\left(\frac{\gamma}{\alpha}\right)^{q-j} p_q \tag{15}$$

At $j = 1$,

$$p_0 = n^q\left(\frac{\gamma}{\alpha}\right)^q p_q \tag{16}$$

Now solve 9 through 16 to yield the proportions of time in terms of p_0, the proportion of time with no vacancies or workers in the queue.

$$p_{-j} = \frac{n!}{(n-j)!}\left(\frac{\gamma}{\alpha}\right)^j p_0,\ j = 1, n \tag{17}$$

$$p_j = n^{-j}\left(\frac{\alpha}{\gamma}\right)^j p_0,\ j = 1, q \tag{18}$$

The sum of all the proportions of time must add up to 1, so that

$$p_0 = \left(\sum_{j=1}^{n}\frac{n!}{(n-j)!}\left(\frac{\gamma}{\alpha}\right)^j + 1 + \sum_{j=1}^{q} n^{-j}\left(\frac{\alpha}{\gamma}\right)^j\right)^{-1} \tag{19}$$

An immediate consequence of 17 through 19 is that it is possible to calculate the proportions of time that the firm has vacancies or workers in the queue. The effects of variables and parameters on these proportions follow directly. The proportion of time the firm has vacancies decreases and the proportion of time the firm has workers in the queue increases as the arrival rate α increases, the departure rate γ decreases, or the balk level q increases.

2.3 Balk Level q

Workers are free to balk at joining a queue to get a job. If wage and arrival rates were the same for each firm, a worker would balk at joining a queue if the expected wait time until employed would be greater than if the worker continued searching for a vacancy or a shorter queue. The comparison generates a condition that determines the optimal balk level for workers in a symmetric equilibrium.

Consider first the expected wait until getting a job when there are j workers currently in the queue and the balk level among workers is q, $j < q$. The rate at which the firm experiences the loss of a worker is γn. If the worker joins the queue, $j + 1$ workers must leave the firm before the worker gets a job. The probability density function for the waiting time until $j + 1$ events is

$$f_j[t] = \frac{\gamma n \left(\gamma n t\right)^j e^{-\gamma n t}}{j!} \tag{20}$$

The expected time until the worker gets a job, if there are already j workers in the queue, is

$$\frac{j+1}{\gamma n} \tag{21}$$

If there are $q - 1$ workers in the queue, and the worker joins the queue, the expected wait would then be $q/(\gamma n)$.

Now consider the alternative that arises if the worker continues searching. Suppose the worker meets firms at the rate m, a constant that does not depend on the number of workers searching.[6] The worker gets a job immediately if the firm has a vacancy, joins the queue if the length is less than

[6] As in search theory, it would be possible to incorporate a matching function relating the rate of meetings between workers and firms to the ratio of searching workers to firms. At this stage in the development of labor queues, the simpler assumption that m is constant is adopted.

the worker's balk level, and continues searching if the queue length equals or exceeds the worker's balk level. In a symmetric equilibrium, all workers would have the same balk level and all firms would have the same proportions of time with vacancies and queues. The expected proportion of time a firm is in a particular state will equal the expected proportion of all firms in that state at a particular point of time. If the balk level for all workers and firms is q, the expected wait time until employment, conditional on a firm having a queue to enter, is

$$\frac{\sum_{j=0}^{q-1} p_j (j+1)/(\gamma n)}{\sum_{j=0}^{q-1} p_j} \tag{22}$$

The expected time until employment for a searching worker, $E_S[q]$, satisfies the implicit condition generated by the alternatives facing a searching worker:

$$E_S[q] = \frac{1}{m} + 0 \sum_{j=1}^{n} p_{-j} + \sum_{j=0}^{q-1} p_j \frac{\sum_{j=0}^{q-1} p_j (j+1)/(\gamma n)}{\sum_{j=0}^{q-1} p_j} + p_q E_S[q] \tag{23}$$

Solving for $E_s[q]$ yields

$$E_s[q] = \frac{1}{1-p_q} \left(\frac{1}{m} + \sum_{j=0}^{q-1} p_j (j+1)/(\gamma n) \right) \tag{24}$$

Now consider whether q is a balk level that could prevail for all workers in equilibrium. If

$$\frac{q}{\gamma n} \leq E_S[q] \leq \frac{q+1}{\gamma n} \tag{25}$$

then workers facing queue length $q-1$ or less would join the queue. Additionally, workers would balk at joining a queue of length q. Then q would be an equilibrium balk level. Suppose instead that $q/(\gamma n)$ is also less than $E_S[q]$. Then consider $q+1$ as the possible equilibrium balk level. In recalculating the expected wait time from searching for balk level $q+1$, $E_S[q+1]$ could increase or decrease. If $E_S[q+1] > q/(\gamma n)$, $q+1$ could then be considered as the equilibrium balk level. Suppose instead that $E_S[q+1]$ decreases (i.e., is less than $E_S[q]$) and $E_S[q+1] < q/(\gamma n)$ if all workers chose balk level $q+1$. Then there would be a mix of workers choosing balk levels q and $q+1$ such that workers would be indifferent between the two balk levels and the expected time until employment would be $q/(\gamma n)$ for workers with both balk levels.[7]. Considering sequentially higher balk levels (and checking for mixed balk levels), a balk level consistent with symmetric equilibrium arises for the first value such that 25 holds.

Figure 2 shows the calculation of the balk level for particular parametric assumptions, assuming workers face a single arrival rate in a symmetric equilibrium.[8] It exhibits general features of the relation between the balk level and the arrival rate. Since the balk level only takes integer values, it is a non-decreasing step function of the arrival rate. At low arrival rates, firms have many vacancies on average and the expected time until employment depends mostly on the rate at which searching workers meet with firms. Then the balk level will be constant over a wide range at low arrival rates, as shown in the figure.[9] As the arrival rate increases, vacancies at firms decline, the expected wait until employment for searching workers increases, and the optimal balk level goes up. At high arrival rates, firms generally have queues and few vacancies, and searching workers must find a queue to join in order to get a job. The balk level then goes up rapidly as vacancies disappear.

The rapid increase in the balk level at higher arrival rates can also be understood in terms of formal queueing theory. Consider a queueing process with arrival rate α, service rate σ at a single server, and no balking. As the ratio α/σ (the erlang or traffic intensity) increases and approaches one, the expected queue length increases indefinitely. In labor queues, as the ratio $\alpha/(\gamma n)$ increases, the expected queue length also increases. With reduced alternatives to joining a queue as $\alpha/(\gamma n)$ increases, workers increase the length of the queue they are willing to join.

2.4 Worker Times Spent in Each State

As a step towards the construction of asset values for workers in each state, it is necessary to calculate the proportions of time workers spend in each state, taking the arrival and departure rates and the balk level as given. Let T_i be the proportion of time workers spend in state i, where the states are

[7] The specifics of such a solution can be determined from the condition that the expected times until employment equal $q/(\gamma n)$.

[8] The figure assumes $n = 20$, $p = 1$, $b = .2$, $\gamma = .1$, $r = .05$, $m = 1$, and $K = 2$. Variation in arrival rate arises from variation in the ratio of workers to firms.

[9] If instead the meeting rate depended on the ratio of vacancies to searching workers, as in the standard matching function for search models, the expected time until employment would approach zero at lower arrival rates and the balk level would also approach zero.

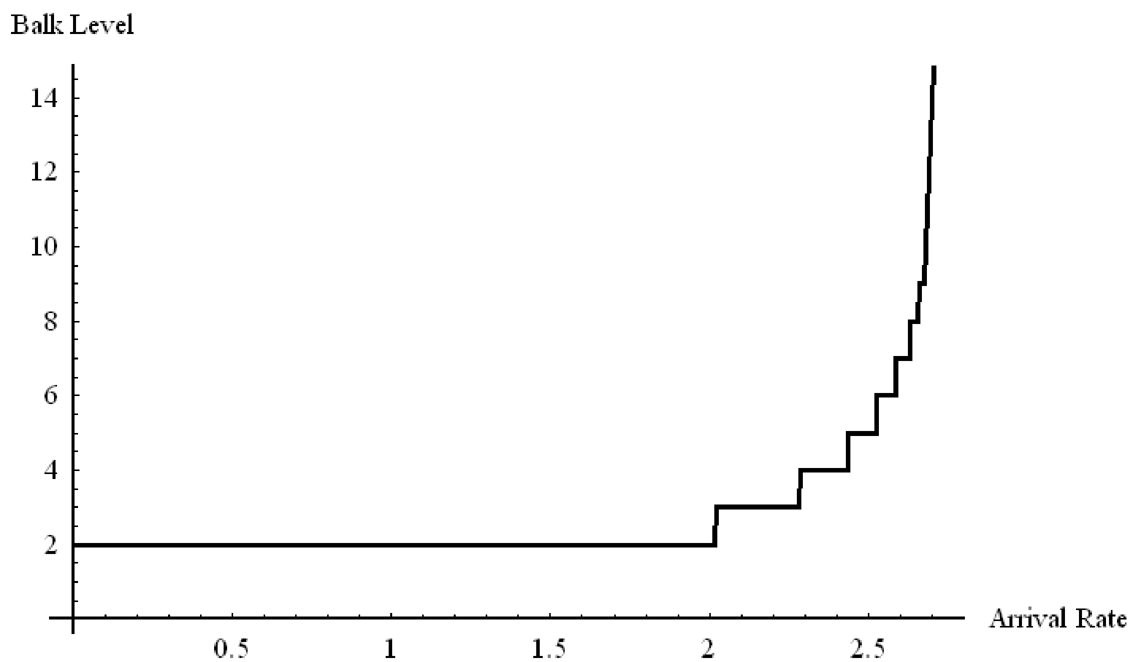

Figure 2: Balk Level versus Arrival Rate

E, Q and S for employed, queueing and searching, respectively. Let S_{ij} be the transition rate for workers from state i to state j, where the possible states are again E, Q and S. Equating flows of workers into and out of states generates the following three conditions.

$$S_{QE}T_Q + S_{SE}T_S = S_{ES}T_E \tag{26}$$
$$S_{ES}T_E = S_{SE}T_S + S_{SQ}T_S \tag{27}$$
$$S_{SQ}T_S = S_{QE}T_Q \tag{28}$$

In the above,

$$S_{ES} = \gamma \tag{29}$$
$$S_{SE} = m\sum\nolimits_{j=1}^{n} p_{-j} \tag{30}$$
$$S_{SQ} = m\sum\nolimits_{j=0}^{q-1} p_j \tag{31}$$
$$S_{QE} = \gamma n \sum\nolimits_{j=1}^{q} p_j / \sum\nolimits_{j=1}^{q} j p_j \tag{32}$$

In 32, S_{QE} solves the condition that the average number of workers in a queue (given that there is a queue) times the transition rate into employment equals γn, the rate at which jobs become available when there are n workers employed. Imposing the condition that $T_E + T_S + T_Q = 1$ and substituting the transition rates from 29 through 32, it is possible to solve for T_E, T_S and T_Q.

2.5 Arrival Rate α

The arrival rate α is determined endogenously and in a symmetric equilibrium does not depend on the wage rate. Let N_W be the number of workers and let N_F be the number of firms (assuming the labor force is fixed). The expected number of workers at a firm is given by

$$F_E = \sum\nolimits_{j=1}^{n} (n-j)p_{-j} + \sum\nolimits_{j=0}^{q} n p_j \tag{33}$$

In equilibrium, the number of workers employed can be calculated as the number of workers times the proportion of time employed or as the number of firms times expected employment per firm. Setting the two measures of the number of workers employed equal yields:

$$N_W T_E = N_F F_E \tag{34}$$

Since T_E and F_E are functions of α, 34 provides a condition on α (treating the balk level q as being determined by α). As α approaches zero, employment per firm approaches zero, the transition rate from searching to employment approaches m (since almost all firms have vacancies), and T_E approaches $m/(m+\gamma)$. As α increases, employment per firm increases, the balk level goes up or stays the same, and T_E declines. As α continues to increase, employment per firm approaches n and T_E approaches zero. There will therefore exist an arrival rate such that 34 holds, and the arrival rate will be unique. The relation between the arrival rate α and the ratio of workers to firms can be found by solving 34 for N_W/N_F as a function of α.

2.6 Worker Asset Values for Each State

Let V_i, $i = E$, S and Q, be the expected asset values for workers entering each of the three states.[10] Since workers in a queue become employed after some time, V_Q can be calculated from V_E by discounting. Let r be the discount rate used by workers to calculate the present value of future income. Then if a worker enters employment t units of time later, the present value of the asset value V_E is $e^{-rt}V_E$, i.e. V_E is discounted by a factor e^{-rt}. If there are j workers in line, the expected discount factor is

$$\int_0^\infty f_j[x]e^{-rx}dx = \left(\frac{\gamma n}{\gamma n + r}\right)^j \tag{35}$$

where $f_j[t]$ is from 20. Multiplying discount factors by the probabilities of queue lengths and summing over queue lengths yields the expected discount factor, C_{DF} :

$$C_{DF} = \sum\nolimits_{j=0}^{q-1} p_j \left(\frac{\gamma n}{\gamma n + r}\right)^j = \frac{\gamma n - \alpha}{\gamma n - \alpha + r} \frac{1 - \left(\frac{\alpha}{\gamma n + r}\right)^q}{1 - \left(\frac{\alpha}{\gamma n}\right)^q} \tag{36}$$

[10] The expected asset value while in the queue will differ from the expected asset value upon entering a queue since the expected time until employment will decline as the worker advances in the queue.

Then

$$V_Q = C_{DF} V_E \tag{37}$$

The other relations among the asset values are

$$\begin{aligned} rV_E &= w + \gamma(V_S - V_E) & (38) \\ rV_S &= b + S_{SE}(V_E - V_S) + S_{SQ}(V_Q - V_S) & (39) \end{aligned}$$

where S_{SE} and S_{SQ} are given by 30 and 31. Then it is possible to solve for the worker asset values V_S, V_E and V_Q. The solution for V_S is:

$$V_S = \frac{w(bS_{SE} + C_{DF}S_{SQ})}{r(r + bS_{SE} + S_{SQ}) + \gamma(r + (1 - C_{DF})S_{SQ})} \tag{40}$$

where C_{DF}, S_{SE} and S_{SQ} are functions of α.

2.7 Firm Profits

As a further step towards determination of the equilibrium wage rate, firm profits can be expressed as

$$\pi = (p - w)F_E - rnK \tag{41}$$

where p is the price of output, w is the wage rate, F_E is expected firm employment from 33, and K is the amount of capital required per position at the firm. The interest rate r is assumed to be the same as the worker discount rate, although this simplification is not essential. Firms choose a wage rate to maximize firm profits, with F_E depending on the arrival rate and corresponding balk level at the firm.

2.8 Wage Determination

Wage determination in the symmetric equilibrium of a labor queueing market follows the methodology in queueing markets for products (Sattinger, 2002) and is based on similar methods in product and labor markets (Carlton, 1978, 1989; Sattinger, 1990; and Moen, 1997). The method assumes that a firm can attract more workers to search at it by offering a higher wage, as in models of directed search with wage posting.

Suppose that workers are able to achieve an expected value of searching given by $\overline{V}_S$. Each firm is small relative to the labor market and takes $\overline{V}_S$ as given. Suppose that, starting from a symmetric equilibrium, a firm deviates by offering a higher wage than other firms. If workers can influence where they search, they would initially be more likely to search at the firm with the higher wage. To be consistent with the previous assumption that workers get interviews at the rate m, assume that workers can change the probabilities that interviews occur at specific firms without changing the interview rate. Then in response to a higher wage rate, the arrival rate at the firm would increase, reducing the likelihood of vacancies, lengthening the expected queue length, and reducing the value of searching at the firm. Workers could choose an optimal balk level at the firm that could differ from balk levels at other firms. Eventually, the arrival rate at the firm would rise to a level such that the value of searching at the firm, given by V_S in 40, falls to the level in the rest of the labor market, $\overline{V_S}$. At that arrival rate, workers would not change further their probabilities of getting interviews at different firms. In response to a wage rate that deviates from the prevailing wage rate, worker behavior generates an arrival rate such that the value of searching at the firm is the same as searching at other firms in the labor market. Treating the arrival rate at the firm α_i as a function of the firm's wage rate w_i, the firm maximizes profits in 41 with respect to w_i. The second order condition on maximizing firm profits requires

$$(p - w_i)\left(\frac{\partial^2 F_E/\partial\alpha_i^2}{\partial F_E/\partial\alpha_i} - \frac{\partial^2\alpha_i/\partial w_i^2}{\partial\alpha_i/\partial w_i}\right) - 2 < 0 \tag{42}$$

The symmetric equilibrium for the labor market consists of an arrival rate satisfying 34, and a value of searching $\overline{V_S}$ and wage rate w that satisfy 37 through 39 and the firm's first and second order conditions for maximization of profits (with firm proportions and balk level determined as before from α).

Figure 3 shows how wages are determined in a symmetric equilibrium, using the same parametric assumptions as in Figure 2 and an equilibrium arrival rate of 1.6.[11] The equilibrium solution (with $\alpha = 1.6$ and $w = .899$) yields a value of searching in the market of $\overline{V_S} = 15.93$. The indifference curve in Figure 3 shows all combinations of wage and arrival rate at a firm that yield $\overline{V_S}$. The equilibrium also generates firm profits of -.446. The isoprofit

[11] Moen (1997, p. 393) provides an analogous figure showing the determination of the wage.

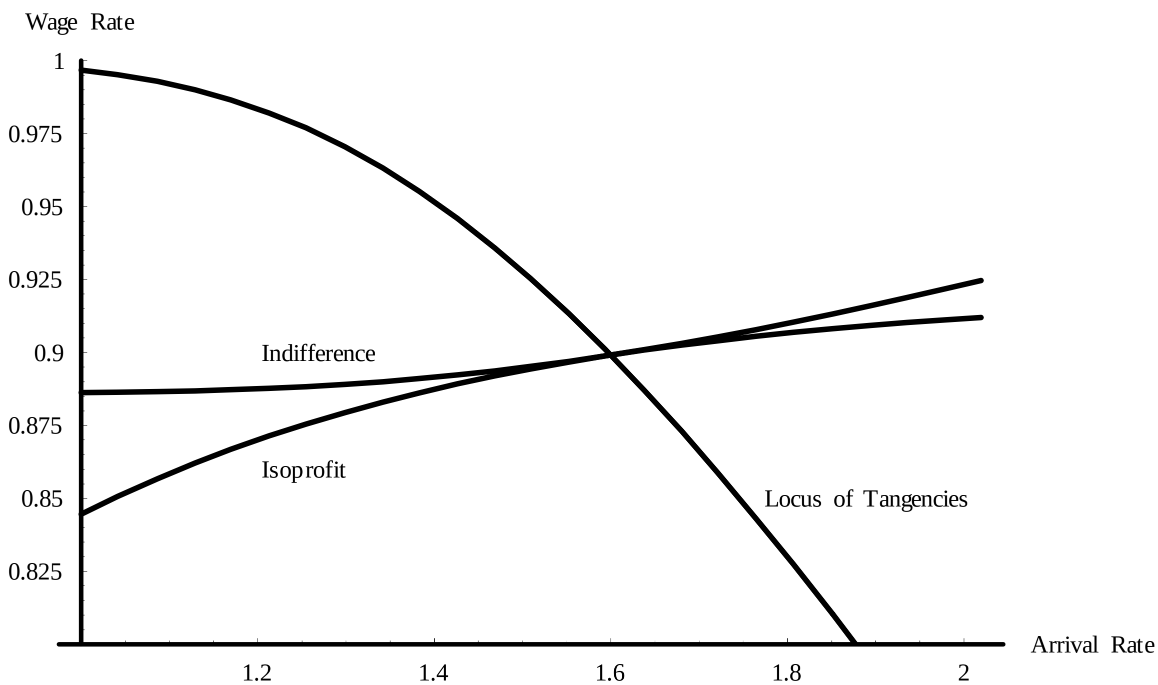

Figure 3: Wage Determination

curve shows all combinations of wage and arrival rates that yield the same level of profits. At the point of tangency, firms are maximizing their profits, taking $\overline{V_S}$ as given. The locus of tangencies in Figure 3 shows the equilibria that would arise with other arrival rates.

3 Dual Wage Equilibria at Higher Balk Levels

3.1 Incentives to Deviate from a Symmetric Solution

The previous section described the symmetric equilibrium that would prevail at arrival rates such that the balk level was at a minimum. To understand why a symmetric equilibrium would not prevail at higher arrival rates, suppose all firms faced an arrival rate $\overline{\alpha}$ that yields a higher balk level than the minimum and consider a firm's incentive to deviate from that solution. To be specific, suppose the parametric assumptions for Figure 1 hold and suppose all firms face an arrival rate of $\overline{\alpha} = 2.2$. Then workers would choose a balk level $q = 3$, higher than the minimum balk level of 2 that holds for arrival rates between 0 and 2.019.

At $\overline{\alpha} = 2.2$, the cost to workers of the increase in balk level to 3 is greater than the benefit of the higher balk level to firms. Specifically, if all workers could agree to set $q = 2$, workers would be just as well off if the wage rate were .565, but firms would be just as well off if the wage were .567. Although a solution with $q = 2$ and a wage rate of .566 would make both workers and firms better off, this solution is infeasible since it is inconsistent with individual worker choice of balk level.

With the parameter values from Figure 2, an individual firm has an incentive to deviate from the symmetric solution by lowering the wage a small amount. Define α_{qj} as the arrival rate at which the balk level jumps from j to $j + 1$. In response to the lower wage rate, workers would refuse to search at the firm as long as the arrival rate were greater than α_{q2}. When the arrival rate falls to α_{q2}, workers would be willing to search at the firm but would set a balk level of 2 (workers could do better by continuing to search than by being third in the queue at a firm paying a lower wage rate than other firms in the labor market). By offering workers a lower balk level and arrival rate, a firm can lower the wage enough to make a greater profit than at the symmetric solution. Therefore when $\alpha > \alpha_{q2}$, the symmetric solution cannot be an equilibrium.

The condition for firms to deviate from a symmetric solution at $\overline{\alpha}$ can be developed formally as follows. Let $\overline{V_S}$ be the value to workers of searching at firms in the symmetric solution at $\overline{\alpha}$, and let $\overline{P_R}$ be the profits of those firms. Let $W_W(V_S, \alpha, q)$ be the wage rate for a worker that would yield a value of searching V_S at a firm with arrival rate α if the balk level is q. Let $W_F(P_R, \alpha, q)$ be the wage rate for a firm that would yield a profit rate P_R when the arrival rate is α and the balk level is q. Suppose j is less than the balk rate at $\overline{\alpha}$. Then a firm has an incentive to deviate from a symmetric solution at $\overline{\alpha}$ whenever

$$W_F(\overline{P_R}, \alpha_{qj}, j) - W_W(\overline{V_S}, \alpha_{qj}, j) > 0 \tag{43}$$

The reason firms have an incentive to deviate from a symmetric solution is that worker choice of a higher balk level generates more negative externalities for other workers than it generates positive externalities for firms. Firms then have an opportunity to offer a better combination of labor market opportunities at a lower wage.

3.2 Equilibrium with Higher Arrival Rates

Now consider how the labor market evolves as more firms start to offer lower wages for lower balk rates. As the number of firms offering a lower wage at α_{q2} increases, the arrival rate at remaining firms increases. Then there are two possible types of equilibrium in the market. First, suppose $\overline{\alpha}$ reaches a level (less than α_{q3}) such that the symmetric solution generated for $\overline{\alpha}$ in Section 2 yields

$$W_F(\overline{P_R}, \alpha_{q2}, 2) - W_W(\overline{V_S}, \alpha_{q2}, 2) = 0 \tag{44}$$

Then at $w = W_F(\overline{P_R}, \alpha_{q2}, 2)$, firms would make the same profits whether they chose α_{q2} or $\overline{\alpha}$, and workers would be indifferent between searching at the two types of firms. (The numbers of workers and firms choosing the two submarkets will be determined in the next subsection.) Since workers and firms are indifferent between the two submarkets, the wage rate, value of searching and profit rate for $\overline{\alpha}$ will be as determined in Section 2 for a symmetric equilibrium.[12] Considering the two markets together, there will be two arrival rates, two unemployment rates and two wage rates. For firms choosing $\overline{\alpha}$, the wage rate and unemployment rate will be higher than for firms choosing α_{q2}.

If 44 does not hold for some $\overline{\alpha}$ between α_{q2} and α_{q3}, then firms would continue to prefer α_{q2} until $\overline{\alpha} = \alpha_{q3}$. Then the wage rate would no longer be determined by the tangency condition in the symmetric solution of Section 2. Instead, there would be two wage rates, w_{q2} and w_{q3}, such that the worker values of searching would be the same at firms with the two arrival rates, and firm profits would also be the same. These two conditions (equality of the values of searching and of the profit rates) generate analytical expressions for the two wage rates, so the solution exists and is unique. The wage rate is higher at firms with the higher arrival rate and balk level. The higher wage rate compensates for a lower expected time employed. The wage rates do not depend on the proportions of firms choosing α_{q2} and α_{q3}.

Since α_{q3} is a corner solution for firms (the maximum arrival rate without workers increasing their balk levels to 3), the condition for equilibrium is that the slope of the isoprofit curve (the change in the wage rate in response to a change in the arrival rate, holding the balk level fixed) should be steeper than the indifference curve at α_{q3} and w_{q3}. Then firms would not gain by offering a marginally lower arrival rate (by lowering the wage) and could not offer a higher arrival rate without generating a higher balk level that would make workers much worse off.

3.3 Proportions of Firms and Workers Choosing Arrival Rate Submarkets

If the equilibrium consists of some firms choosing one arrival rate and the rest choosing a higher arrival rate (with a higher balk level), the proportions of workers and firms choosing each can be determined in a straightforward manner. Suppose that the two arrival rates are α_1 and α_2, with the average rate at all firms given by α, $\alpha_1 < \alpha < \alpha_2$. Let P_{F1} be the proportion of firms choosing α_1. Then

$$P_{F1}\alpha_1 + (1 - P_{F1})\alpha_2 = \alpha \tag{45}$$

[12] Assume workers and firms consistently choose one type of submarket over time. Then the submarket with arrival rate $\overline{\alpha}$ can be considered in isolation, and the symmetric equilibrium from Section 2 applies.

Rearranging,

$$P_{F1} = \frac{\alpha_2 - \alpha}{\alpha_2 - \alpha_1} \tag{46}$$

Let P_{W1} be the proportion of workers choosing α_1, let $(N_W/N_F)_i$ be the ratio of workers to firms choosing α_i, and let N_W/N_F be the ratio of all workers to all firms for the aggregate labor market. Then

$$(N_W/N_F)_1 = \frac{P_{W1}}{P_{F1}} \frac{N_W}{N_F} \tag{47}$$

$$(N_W/N_F)_2 = \frac{(1 - P_{W1})}{(1 - P_{F1})} \frac{N_W}{N_F} \tag{48}$$

From 34,

$$(N_W/N_F)_i = \frac{F_{\text{Ei}}}{T_{\text{Ei}}} \tag{49}$$

where F_{Ei} is employment per firm with arrival rate α_i from 33 and T_{Ei} is proportion of worker time employed from Section 2.4. Then it is possible to solve 47 and 48 for the proportion of workers P_{W1} and aggregate ratio N_W/N_F in terms of α, α_1 and α_2. The worker proportion P_{W1} approaches zero as P_{F1} approaches zero, and approaches 1 as P_{F1} approaches 1. The relationship is roughly (but not exactly) linear.

3.4 Dual Wage Numerical Solutions

This section calculates the dual wage equilibria for arrival rates that generate balk rates greater than the minimum, using the parametric assumptions for Figure 2. Table 1 presents the arrival rates at which balk levels jump, together with tests for the presence of a dual wage equilibrium. In the row for balk level 2, the lower arrival rate is α_{q2}, the arrival rate at which the balk level jumps to 3. The upper arrival rate, in column 3, is the arrival rate at which the balk level jumps to the next level, 4. The incentive to deviate in column 4 presents the calculation of $W_F(\overline{P_R}, \alpha_{q2}, 2) - W_W(\overline{V_S}, \alpha_{q2}, 2)$. When this difference is positive, firms would have an incentive to offer a lower wage and a lower balk rate for some interval of arrival rates above α_{q2}. The condition for a corner solution in column 5 calculates the ratio of the slope of isoprofit curve to the slope of the indifference curve at α_{q3}. When this ratio is greater than one, firms with arrival rate α_{q3} and balk level 3 will not have an incentive to raise or lower the arrival rate. Table 2 shows the conditions at the lower and upper arrival rates for dual wage equilibria, including workers per firm, unemployment rate and wage rate.

4 Vacancies, Unemployment, Queueing and Search

This section demonstrates the underlying relations derived in labor queues. A fundamental result of search theory is that there will be an inverse relation between firm vacancies and worker unemployment. This relation is generated by the matching function, which yields the transition rates for unemployed workers and firms with vacancies as a function of the ratio of vacancies to unemployed (market tightness). Analogously, the results in the previous sections on labor queues generate an inverse relation between firm vacancies and worker unemployment as a consequence of the proportion of time a firm has vacancies or queues shorter than the balk level. With labor queues, however, unemployment consists of workers either searching or waiting in queues. Labor market conditions (as reflected in the firm arrival rate α) determine the mix of unemployed workers between those searching and queueing. As α increases, the labor market passes from circumstances where most unemployed workers are searching to circumstances where most are in queues.

4.1 Arrival Rate Versus Workers per Firm

From Section 2.2, employment per firm, F_E, is an increasing function of the arrival rate α, and the proportion of worker time employed, T_E, is a decreasing function (with the balk level q optimal at each α). In the dual wage equilibria, the results of section 3.3 yield a relation between the average arrival rate α and the aggregate ratio of workers to firms, $N_W/N_{F.}$ Combining the results from the symmetric and dual wage equilibria, the equilibrium arrival rate will be a non-decreasing function of the ratio of workers per firm. Figure 4 shows the relationship for the same parametric assumptions as Figure 2 and exhibits some general features.

At low ratios of workers per firm, most firms have vacancies because of the low arrival rate, worker expected time to employment is constant, and the balk level is also constant at the minimum level. When vacancies decline at higher ratios of workers per firm, workers would raise their balk level as

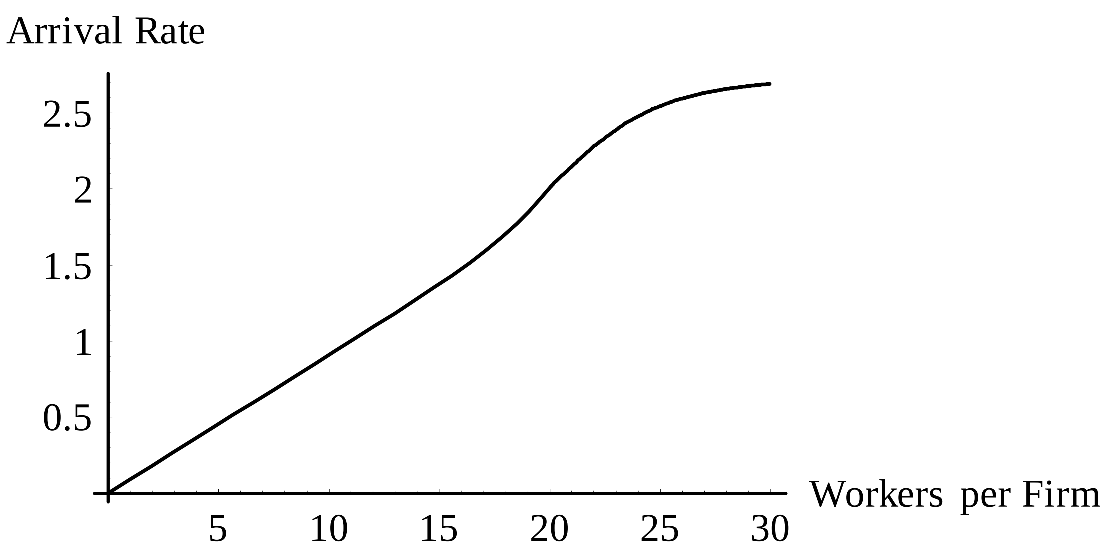

Figure 4: Average Arrival Rate per Firm

described in the previous sections. At the arrival rate at which a balk level increases from the minimum, additional workers (from more workers per firm) are absorbed in queues in the dual wage equilibrium. As the ratio of workers per firm reaches high levels, vacancies decline to zero, and additional workers are absorbed in increasingly lengthy queues without raising the arrival rate significantly. As a consequence, the arrival rate approaches a maximal level.

4.2 Proportions of Workers Searching and Queueing

With queues, the labor market is qualitatively different at low and high ratios of workers per firm. At a low ratio of workers per firm, firms often have vacancies and seldom have queues. Unemployed workers are essentially searching for vacant positions as in standard search theory. At a high ratio of workers to firms, the arrival rate is high and firms rarely have vacancies. Workers, to get a job, must find a firm with a queue shorter than their balk level. Most unemployed workers are then in queues rather than actively searching. Proportion of time spent searching remains substantially constant, however, because workers must continue searching when they meet firms with queues at the balk level.

Figure 5 shows proportions of workers searching and queueing as functions of the ratio of workers to firms. The figure is constructed by combining results from the symmetric equilibrium (in the interval for workers per firm generating the minimum balk level) with the results for dual wage equilibria at higher intervals of workers per firm. Proportion of time queueing starts at negligible levels for low ratios of workers per firm and increases rapidly at high levels when vacancies disappear. The proportion of workers searching varies over a small range when workers per firm increase, as workers shift from searching for vacancies to searching for queues shorter than their balk levels.

4.3 Unemployment Versus Vacancies

A fundamental accomplishment of search theory is the generation of the relation between unemployment and vacancies observed by Beveridge. This relation is constructed by deriving the transition rates between employment and unemployment and between vacant and filled jobs from the matching function. Queueing provides an alternative relation between unemployment and vacancy rates generated by the proportion of a firm's positions that are

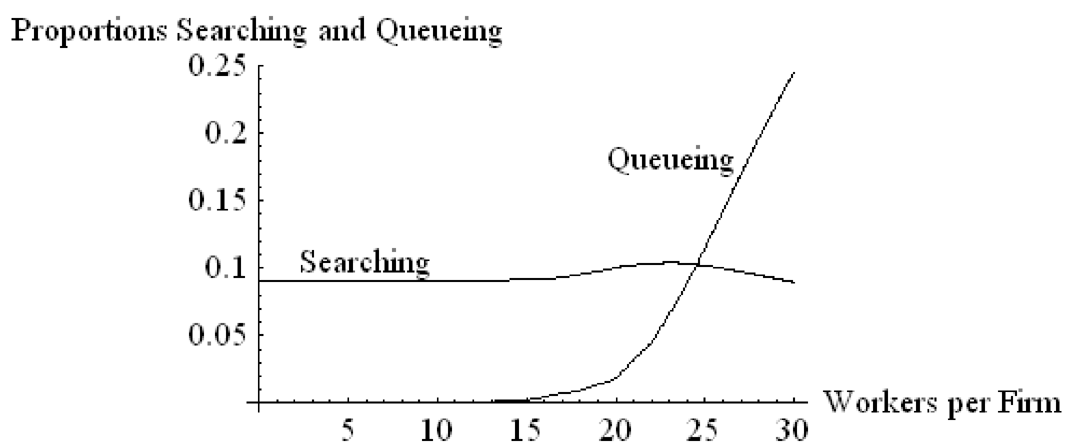

Figure 5: Proportions Searching and Queueing

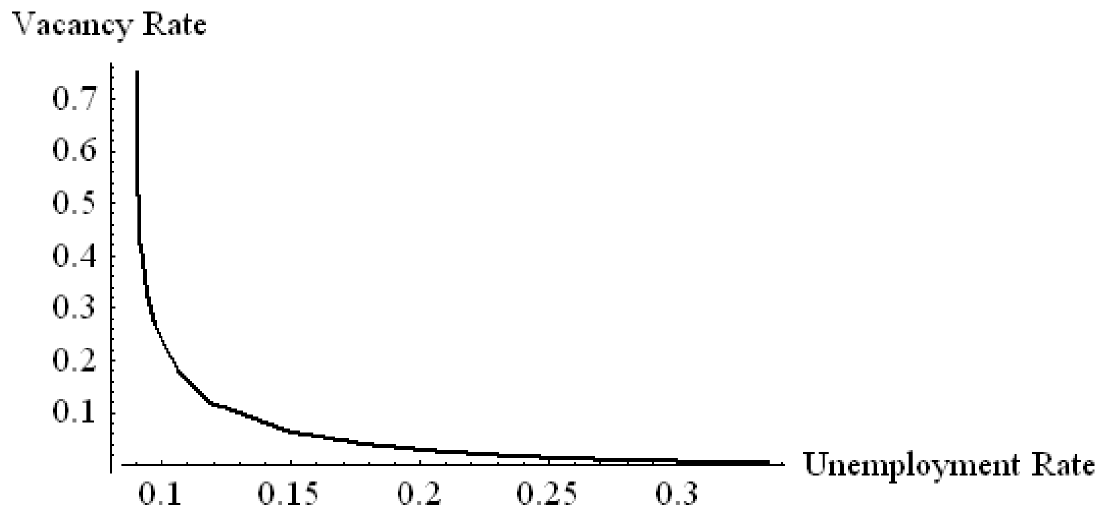

Figure 6: Beveridge Curve

vacant and the proportion of time workers are not employed. Unemployment in a queueing model of the labor market is defined here as including searching as well as queueing workers.[13]

Figure 6 shows the Beveridge Curve generated by queueing using the same assumptions as for earlier figures. Different points on the Beveridge Curve arise from different ratios of workers to firms and the steady states resulting from arrival rates. As with Figures 3 and 4, it combines results from the symmetric and dual wage equilibria. The vacancy rate is given by $1 - F_E/n$, and the unemployment rate (combining searching and queueing) is given by $1 - T_E$. The curve shares the general inverse shape with the Beveridge Curve derived from standard search theory.

[13] Bureau of Labor Statistics procedures may not classify queueing workers as unemployed. See Section 5.

5 Conclusions

This paper has applied methods of queueing theory to the analysis of a labor market with frictions. Queueing theory provides expressions for the equilibrium numbers of vacancies, searching workers and queueing workers in terms of the arrival rate. As the ratio of workers to firms and the arrival rate increase, unemployed workers shift from searching to queueing and the balk level increases. At arrival rates that would generate a balk level greater than the minimum, a dual wage equilibrium arises in which some firms have an arrival rate at which the balk level would jump above the minimum level while other firms would have a higher arrival rate (and a higher balk level and wage rate). The results generate a Beveridge Curve without the assumption of a non-trivial matching function. Queueing theory also carries implications for the dynamic adjustment of labor markets with frictions. Changes in unemployment take place mostly through the number of workers in queues rather than through the number of workers actively searching. In a comparison with the standard search and matching model, queueing places limits on the numbers of unemployed workers who are searching. As a result, at higher ratios of workers per firm, the arrival rate is lower when there is queueing.

This paper only sets forth a basic model of queueing in a labor market. With this model established, there are several immediate questions that can be addressed. First, how do outcomes of the model differ from outcomes when queueing is absent? In particular, how is the ratio of vacancies to unemployed affected by queueing? An evaluation of labor queues requires that models with and without queues be placed on a comparable basis, so that the only difference is the presence or absence of queues. One immediate difference between the labor queues model introduced in Section 2 and the standard search and matching model (Pissarides, 2000; Mortensen and Pissarides, 1999) is the use of a matching function. Although the labor queueing model can be constructed without a matching function, there is no inconsistency and a matching function can be incorporated to facilitate comparison. A second difference is that in the standard search and matching model, a firm is coextensive with a single job. The job moves between the two states of filled and vacant, generating a well-defined vacancy. Despite the almost universal adoption of this simplifying assumption, it is unnecessary for the construction of the basic relationships in a search and matching model. Section 2 describes methods that can be applied to determine the probabilities that a firm with up to n workers has 0 through n vacancies. In the standard model, an unemployed worker searches only among jobs that are vacant. The number of matches then depends on the numbers of unemployed workers and vacant jobs. When firms are greater than a single job, it is necessary to specify how the arrival rate varies with the number of vacancies at the firm. However, to be consistent with the determination of state probabilities in Section 2, the arrival rate must be independent of the number of vacancies or number in queue. One resolution is to assume searching workers seek jobs or queues at all firms, not just firms with vacancies. Specifically, assume the number of interviews between workers and firms is given by $m[S, N_f]$, where S is the number of searching workers not in queues, and N_f is the number of firms.[14] An interview results in employment if the firm has a vacancy. If there are queues and the queue length is less than the balk level, the interview could result in a transition of the worker from searching to queueing. As in the standard literature, assume $m[S, N_f]$ is a continuous, concave function of its arguments and is homogeneous of degree one. With these assumptions, the same matching function can be used for the model with and without queueing. In this extension, queueing again places limits on the number of unemployed workers who are searching. As a result, at higher ratios of workers per firm, the arrival rate is lower when there is queueing. Then the wage rate is less sensitive to the ratio of workers per firms when there is queueing.

Another general area for extensions concerns heterogeneous workers or firms, corresponding to heterogeneous servers or customers. With this extension, the queueing model could explain the assignment of heterogeneous workers to heterogeneous firms. Shimer (2005) considers queueing with heterogeneous workers in the context of coordination frictions. Heterogeneous firms could generate common resource externalities in a queueing model (possibly relevant to the Lagos model, 2000). One source of firm heterogeneity is firm size. Firm size effects have been extensively studied, and queueing theory may provide alternative explanations for observed phenomena.

There are a number of efficiency questions related to queueing. Efficiency of search has been analyzed in terms of search congestion and the externalities associated with the decision to enter the labor market. Search congestion may be affected by queueing since queueing workers reduce the ratio of searching workers to firms. Corresponding to search congestion, queueing theory suggests that there will be externalities from the decision to enter a

[14] This implicitly assumes that firms are passive in the matching process, except for their influence on the arrival rate through the wage rate.

queue because entry increases wait times for those who enter later (Hassin and Haviv, 2003, pp. 18-19). Firm decisions to increase or decrease their queue lengths will affect the arrival rate of workers at other firms. As noted above, heterogeneous firms could generate common resource externalities if workers choose firms on the basis of average queue length rather than marginal queue length increase.

Immediate evidence of labor queues is lacking because of the absence of direct measures in U.S. labor force statistics. In the Current Population Survey, a person who is waiting to start a job is not separately classified as in a queue. Instead, the individual could be classified as unemployed or out of the labor force depending on whether he or she meets the criteria for unemployment (primarily having actively looked for a job within the last four weeks). Before 1994, a person waiting to start a new job within 30 days would be treated as unemployed without regard to their search activity. A consequence of these practices is that if workers queue for jobs, they would be counted as out of the labor force if they do not continue searching. At times of high unemployment, the amount of unemployment would be understated by the numbers of workers in queues. Dynamic changes in the numbers of workers in queues would appear as movements of workers into or out of the labor force. It is not yet clear which labor force data would be used to establish whether workers queue for jobs.

Given the externalities associated with queueing, it would be important to determine whether queueing can increase the efficiency of operation of labor markets. In the comparison using a matching function described above, a labor market with queueing is less efficient than a labor market without queueing. However, when queueing length is limited to one, a labor market with queueing is more efficient at high ratios of workers to firms. Then queueing carries the potential to increase the efficiency of operation of labor markets, by reducing vacancies or unemployment. Queueing may not arise if firms cannot commit to hiring workers in the order in which they queue, so that workers would be unwilling to wait unemployed in a queue for a job. Then a change in labor market institutions (for example, by formalizing the firm's commitment to hire workers in the queue) could bring about an increase in efficiency.

References

[1] Allon, G. and Federgruen, A.: "Competition in Service Industries," unpublished paper, 2004, Columbia University.

[2] Brockmeyer, H.L., Halstrom, and Jensen, Arne: "The Life and Works of A.K. Erlang," *Transactions of the Danish Academy of Technical Sciences*, No. 2, 1948.

[3] Carlton, D. W.: "Market Behavior with Demand Uncertainty and Price Inflexibility." *American Economic Review*, Vol. 68, No. 4, 1978, pp. 571-587.

[4] Carlton, D. W.: "The Theory and the Facts of How Markets Clear: Is Industrial Organization Valuable for Understanding Macroeconomics?" Chapter 15, 909-946, in: Handbook of Industrial Organization, Vol. I, edited by R. Schmalensee and R. Willig. Amsterdam: North Holland, 1989.

[5] Cox, D. R. and Smith, W. L.: Queues. New York: John Wiley, 1961.

[6] Feller, W.: An Introduction to Probability Theory and its Applications, Volume I, second edition. New York: John Wiley, 1957.

[7] Gaumont, Damien, Schindler, Martin, and Wright, Randall: "Alternative Theories of Wage Dispersion." *European Economic Review* Vol. 50, 2006, pp. 831-848.

[8] Gnedenko, B. V., and Kovalenko, I.N.: Introduction to Queueing Theory, second edition, translated by Samuel Kotz. Boston: Birkhauser, 1989.

[9] Gross, Donald, and Harris, Carl M.: Fundamentals of Queueing Theory, second edition. New York: John Wiley, 1985.

[10] Hassin, R. and Haviv, M.: To Queue or Not to Queue; Equilibrium Behavior in Queueing Systems. Boston: Kluwer 2003.

[11] Karlin, S.: A First Course in Stochastic Processes. New York: Academic Press, 1966.

[12] Karlin, S. and Taylor, H. M.: A Second Course in Stochastic Processes. New York: Academic Press, 1981.

[13] Knudsen, Niels Chr. "Individual and Social Optimization in a Multi-server Queue with a General Cost-Benefit Structure." *Econometrica* Vol. 40, No. 3, May 1972, pp. 515-528.

[14] Lagos, Ricardo: "An Alternative Approach to Search Frictions." *Journal of Political Economy* Vol. 108, No. 5, October 2000, pp. 851-873.

[15] Leeman, W.A.: "The Reduction of Queues Through the Use of Price." *Operations Research* Vol. 12, 1964, pp. 783-785.

[16] Moen, Espen R.: "Competitive Search Equilibrium." *Journal of Political Economy* Vol. 105, 1997, pp. 385-411.

[17] Moldovanu, B. and Kittsteiner, T.: "Priority Auctions and Queue Disciplines that Depend on Processing Time." *Management Science* Vol. 51, No. 2, February 2005, pp. 236-248.

[18] Mortensen, D., and Pissarides, C.: "New Developments in Models of Search in the Labor Market." In *Handbook of Labor Economics*, Vol. 3B, edited by Orley C. Ashenfelter and David Card, pp. 849-919, North-Holland, Amsterdam, 1999, pp. 2567-2627.

[19] Naor, P.: "The Regulation of Queue Size by Levying Tolls." *Econometrica* Vol. 37, No. 1, January 1969, pp. 15-24.

[20] Pissarides, Christopher, 2000. *Equilibrium Unemployment Theory,* second edition. Cambridge: MIT Press.

[21] Saaty, T. L.: Elements of Queuing Theory. New York: McGraw-Hill, 1961.

[22] Sattinger, M.: "Unemployment, the Market for Interviews, and Wage Determination." *Journal of Political Economy* Vol. 98, No. 2, April 1990, pp. 356-371.

[23] Sattinger, M.: "A Queuing Model of the Market for Access to Trading Partners." *International Economic Review* Vol. 43, 2002, pp. 533-547.

[24] Sattinger, M.: "Price Dispersion and Short Run Equilibrium in a Queueing Model." Unpublished paper, University at Albany, 2002.

[25] Shimer, R.: "The Assignment of Workers to Jobs in an Economy with Coordination Frictions." *Journal of Political Economy* Vol. 113(5), 2005, pp. 998-1025.

[26] Stahl II, D. O.: "Queue-Rationing and Price Dynamics." *Scandinavian Journal of Economics*, Vol. 89, No. 4, 1987, pp. 469-485.

[27] Taylor, H. M. and Karlin, S.: An Introduction to Stochastic Modeling. Orlando: Academic Press 1984.