Wesleyan Economic Working Papers

http://repec.wesleyan.edu/
N°: 2012-007

Neoclassical Models of Imperfectly Competitive Labor Markets

Joyce P. Jacobsen and Gilbert L. Skillman

August, 2012

WESLEYAN UNIVERSITY

Department of Economics
Public Affairs Center
238 Church Street
Middletown, CT 06459-007

Tel: (860) 685-2340
Fax: (860) 685-2301
http://www.wesleyan.edu/econ

Neoclassical Models of Imperfectly Competitive Labor Markets

Joyce P. Jacobsen and Gilbert L. Skillman

August 17, 2012

Draft of Chapter for *Models of Labor Markets*, Bruce Kaufman (ed.), forthcoming

CHAPTER 3: Neoclassical Models of Imperfectly Competitive Labor Markets

The scenario of perfect competition, as augmented by human capital theory, yields a large body of testable hypotheses concerning labor market outcomes. These hypotheses, overviewed in the first chapter of this volume, might all be understood as applications of three more fundamental implications of perfectly competitive equilibrium conditions: *no unemployment or vacancies*, *the law of one wage*, and *compensating wage differentials*. The hypothesis of no equilibrium unemployment or job vacancies (NUV) follows immediately from the assumption that wage rates flexibly adjust to clear markets, so that the only reason for actors not to be engaged as employees or employers is because they choose not to participate in labor markets at going wage rates. The law of one wage (LOW) asserts that employees with equal (marginal) productivities facing identical working conditions receive the same wage rates, and follows from the assumption of frictionless exchange. Complementarily, the hypothesis of compensating wage differentials (CWD) asserts, broadly speaking, that variations in wage rates across workers are explained by either variations in labor productivity arising from costly investments by workers or variations in working conditions that are costly for employers to modify.

These fundamental implications of competitive theory are challenged by both casual observation and careful empirical analysis. The NUV hypothesis conflicts with the appearance of persistently positive unemployment and vacancy rates. The LOW hypothesis is tested by manifestations of labor market "duality" or segmentation (Dickens and Lang 1993), persistent inter- and intraindustry wage differentials (Krueger and Summers 1988, Groshen 1991b) discrimination by ethnicity, gender, or race (Cain 1986; Jacobsen 2007, Chapter 9), and other indications that workers who are seemingly identical from the standpoint of competitive theory

nonetheless receive unequal wages. Indeed, such indications motivated the initial emergence of labor economics as a distinct field (Kerr 1988). Finally, consistent evidence of CWD has been notoriously hard to establish (Dorman 1996, Chapter 4).

Such inconsistencies help motivate the study of models of imperfect competition in labor markets. Consistently with the observation that employers typically set wage levels, the chapter's primary focus will be on two cases of imperfect competition in which equilibrium wage levels are set by firms, *monopsony* and *efficiency wages*. In one sense, these cases reflect contrary views of employers' wage-setting power, in that the former scenario is generally associated with equilibria in which wages are below their corresponding perfectly competitive levels, while the latter are generally associated with supra-competitive wage levels. Later in the chapter, we will also consider the scenario of *bilateral monopoly* or bargaining that also raises the possibility of supra-competitive equilibrium wages.

The cases of monopsony, efficiency wages, and bargaining will be used to study the conditions underlying four major departures from the fundamental predictions of perfectly competitive theory: 1) equilibrium unemployment and vacancies; 2) the "wage curve," which posits an inverse relationship between sectoral wages and unemployment rates; 3) variations in wages and working conditions inconsistent with LOW and CWD, including wage disparities due to race or gender discrimination and "segmentation" of labor markets; and 4) wage rigidity in response to revenue productivity shocks in the labor market. Thus, while we also address normative concerns raised by these theoretical scenarios, our primary focus is on their positive implications. An extensive complementary discussion of the policy implications of labor market imperfections is provided by Boeri and van Ours (2008).

Although the focus is on imperfectly competitive labor markets and the implications of given departures from competitive conditions, the analysis discussed in this chapter remains neoclassical in flavor, in that hypotheses are derived from the consideration of equilibrium outcomes in which individual actors are assumed to choose actions to maximize their individual (expected) payoffs based on exogenously determined preferences and available information, and given the actions chosen by other actors in the relevant markets. This conception corresponds generally to the notion of Nash equilibrium for appropriately defined non-cooperative games.

1. Exchange costs and forms of labor market imperfections[1]

To appreciate the different forms that imperfect competition can take, it is helpful to frame the discussion of imperfectly competitive labor markets in terms of the costs associated with the constituent activities of labor exchange and their implications for market behavior and performance. These activities can be grouped into three categories: *matching*, *negotiation* and *enforcement*. Introducing the realistic premise that individuals incur costs when engaging in these respective aspects of exchange activity can be expected to have distinct implications for the forms and outcomes of market interaction, as can be seen by contrasting scenarios of costly exchange activity with the case of perfectly competitive markets in which all such activities are assumed to be costless.

1.1 *Matching costs and labor market flows*

Matching refers to those activities required for the pairing of potential exchange partners. It subsumes such activities as *search* for potential exchange partners and the exercise of individual *mobility* as necessary to locate where such exchange partners are available. Labor suppliers

engage in search, for example, when they read job ads, attend job fairs, or undergo interviews with prospective employers, and exercise mobility in moving their households to within commuting distance of these employers. Labor buyers also engage in search (posting vacancy notices and co-hosting job fairs, for instance) and may also mobilize in pursuit of more favorable matches, as when firms move plants to areas with lower-cost labor.

Realistically, matching in labor markets demands expenditures of time, effort, or money, and thus incurs costs quite apart from the costs of expending or compensating labor. To see the implications of such costs for labor market functioning, note by way of contrast that wage-taking behavior, a core premise of the perfectly competitive model, depends in part on the assumption of costless matching. To say that a firm takes the going market wage rate as given, for example, is to say that on one hand, the firm can hire as much labor as it pleases at the going wage rate, and on the other, that no labor supplier would elect to work for the firm should it choose to lower its wage offer even a penny below the going rate. Whatever other condition might also be required (e.g., a large number of firms in the market offering homogeneous working conditions), these conditions presume that would-be employees can move immediately and effortlessly between firms (and into and out of the market), and to that extent find alternative employment prospects, or even non-employment, to be effective substitutes.

In contrast, to the extent that mobility is costly, all of these options become imperfect substitutes, and the economic logic supporting the assumption of wage-taking behavior is correspondingly undermined. Viewed in this light, monopsonistic wage-setting power can be understood as a reflection of matching costs incurred by labor suppliers, with the pure case of monopsony emerging when these costs are sufficiently high to preclude effective competition from any other prospective employer. By the same reasoning, matching costs for labor buyers

endow prospective employees with countervailing wage-setting power. As noted in the introduction, we will consider the case of bilateral market power later in this chapter.

In addition to the emergence of wage-setting power on one or both sides of the market, there is another important qualitative consequence of the existence of matching costs. The necessity of costly search for suitable employment matches raises the possibility of unemployment or vacancies as equilibrium phenomena, rather than as just indications that market equilibrium has not yet been achieved. To animate this point, imagine a scenario in which a representative worker engages in sequential search: at a given cost per period, a typical worker gains a (perhaps uncertain) option to draw from a distribution of wage offers, and, if an offer is received, must decide whether to accept the offer or reject it in the expectation of receiving a more favorable subsequent draw. It is not hard to see that the worker might optimally decide to reject the offer in hand if it is sufficiently low and thus remain unemployed for a current period. Similar reasoning can support the existence of equilibrium job vacancies if we imagine that firms face the possibility in each period of receiving either no applications or applications from workers with relatively poor qualifications for the available positions.

There is some necessary ambiguity as to whether search-related unemployment is voluntary or involuntary. The standard account, according to which a worker is involuntarily unemployed if she is willing to work at the going wage rate but unable to find a job, cannot be directly applied insofar as a nondegenerate distribution of wage offers is assumed to exist. Furthermore, individual decisions to remain unemployed are clearly voluntary to the extent wage offers are received but freely rejected.[2] Replacing "the going wage rate" with "the mean wage rate" addresses the conceptual issue but raises difficulties in practical application, since neither

individual *reservation* wage levels nor histories of wage offers are readily observable, and nothing guarantees that reservation wage levels are less than the mean wage.

There is a large literature on the economics of search (McCall and McCall 2007). We will not attempt to summarize this literature, but it provides one of the key analytical foundations for an overarching vision of how labor markets function that provides a comprehensive alternative to the Walrasian model of competitive equilibrium. This approach, which we explore in this chapter, understands equilibrium as the (evolving) harmonization of *flows* of jobs and workers into and out of the market, rather than as a series of complete pairings of available transaction partners, as in a comparative-static Walrasian framework.

This vision of markets as a flow process has gained increasing currency in neoclassical labor economics since the late 1980s (see, for example, Davis, Haltiwanger and Schuh 1990). In the context of labor markets, the two main types of flows are *job flows* and *worker flows*, both of which affect the extent of unemployment and vacancies in given labor markets. Job flows are generated by forces which create new jobs, immediately establishing new vacancies to be filled, and curtail existing ones, altering both the vacancy *rate* (the ratio of existing vacancies to the sum of vacancies and filled positions) and the level of unemployment. On the other side of labor markets, workers flow across three possible positions vis-à-vis the labor market: out of the labor force, unemployed, and employed. Worker flows from employment to unemployment or out of the labor force are occasioned by quits and layoffs, just as new hires represent flows in the opposite direction. Flows into and out of labor markets affect both the level and the rate of unemployment, where the latter is defined as the ratio of total unemployed to the sum of unemployed and employed workers in the market.

A key element in the conception and analysis of labor market flows is the *matching function*, which summarizes the process that yields pairings of prospective employers and employees. The number of matches in a given period is typically represented as an increasing function of the stock of unemployed workers and vacant positions (or, equivalently, the rate at which matches are made is defined as an increasing function of unemployment and vacancy rates, for given stocks of jobs and workers); the notion that matching is costly is captured by the assumption that not every situation of unemployment or vacancy is immediately terminated in a consummated job match.

An analytically complete model of the labor market as a set of interacting flows would thus involve a specification of the stock of prospective labor suppliers (each of whom, at a given time, is either employed, unemployed, or out of the labor force) and of filled and vacant employment positions in addition to the specification of flows of workers into and out of the labor force, flows of newly created and newly destroyed positions, and the matching function that summarizes the process by which unemployed workers are paired with unfilled job openings. In addition, there must be some representation of what happens after prospective employers and employees are matched. Any such characterization will reflect considerations having to do with the other two constituent activities of exchange, negotiation and enforcement, to be discussed below. Finally, the relevant equilibrium concept for this model is that of a *steady state* in which (expected) flows into and out of unemployment and job vacancies balance out.

1.2 *Negotiation costs and setting the terms of labor exchange*

The standard supply-and-demand model of labor market interactions suggests that setting the terms of labor exchange is a rather trivial matter. There are only two variables in play, the wage

level and hours worked, and the former is given by "the market." For any given level of the market wage rate, workers choose the number of hours they would like to work, and firms choose the profit-maximizing hours of labor to employ. The wage rate then somehow adjusts to equate aggregate quantities of labor supplied and demanded.

This simple analytical story starts to unravel once we grant the prospect that there are potentially many dimensions of interest in play in any real-world employment relationship beyond wage levels and hours worked, including the labor effort or intensity expected per time period, the specific range of tasks to be performed by given workers, the working conditions under which labor is to be performed, and the form and manner of labor compensation. Moreover, the desired specification of terms in all of these dimensions may depend on a large number of contingencies affecting production or the terms of trade.

An immediate implication of this complexity is that it is no longer meaningful to represent competitive labor market conditions by *wage*-taking behavior, since even competitive employees and firms might plausibly consider mutually acceptable tradeoffs between the wage level and other aspects of the employment relationship. In this context, it is more appropriate to assume that competition establishes that the expected *payoff* received by players on one or the other side of the market is given—that is, that firms receive zero economic profit in competitive equilibrium, or workers receive a given reservation utility level. Indeed, this is the foundation on which the analysis of compensating wage differentials is constructed.

Beyond compelling a more elaborate representation of typical labor exchanges, this real-world complexity may have implications for the manner in which the terms of labor exchange are determined. Let *negotiation* connote the process resulting in the specification of terms of trade that are mutually acceptable to all parties of a prospective employment match. To the

extent that there are multiple payoff-relevant dimensions to labor exchange, and that the appropriate specification of the terms of exchange are dependent on the realization of subsequent randomly determined contingencies, arriving at mutually acceptable terms of exchange may itself be a costly enterprise.

For firms that sell many different products or hire many different types of labor, negotiation costs have been put forward to explain the presence of nominal rigidities in prices and wages. When a firm must post its wage schedule in advance (perhaps for reasons suggested in the previous paragraph), it incurs *menu costs* whenever it changes this schedule in response to changing market conditions. To economize on such costs, it may optimally maintain a given nominal wage schedule in the face of changes in market conditions or the overall price level, leading to (short-run) conditions in which equilibrium is achieved through adjustments in quantities rather than relative wages (Akerlof and Yellen 1985).

A second manifestation of negotiation costs, more immediately relevant to our purposes in this chapter, becomes salient when both parties to a prospective employment match face significant costs of rematching with other exchange partners. As anticipated earlier, this gives rise to a situation of bilateral monopoly in which both parties have potential power in determining the division of (quasi-)rents. Since the threat of replacement is not sufficient to drive either party to his or her reservation payoff level, there is scope for the parties to haggle over the surplus generated by the relationship. But this haggling requires the expenditure of time and perhaps other costly resources. In this case, one might expect that the equilibrium surplus division depends on the *relative* negotiation costs of the two parties, with the party enjoying a cost advantage having greater bargaining power. The bargaining implications of negotiation and mobility costs will be discussed later in the chapter.

Bargaining and negotiation models of wage setting play an important role in the study of labor market scenarios in which the presence of bilateral power is central to the set of problems under study. We consider these scenarios later in the chapter.

1.3 *Enforcement costs and incentive provisions*

The competitive model of labor exchange is based on the premise that the terms of exchange, once agreed to, are perfectly enforceable at no cost. This scenario presumes, in effect, that all of the relevant terms of labor exchange can be fully specified in a contract and completely enforced, all at no cost to the transacting parties. In this ideal contracting scenario, externally applied legal penalties suffice to ensure that the negotiated terms of trade will be carried out, once agreed to.

Certain aspects of real-world labor exchanges challenge the relevance of the ideal contracting scenario. On one hand, it may be prohibitively costly to identify and contractually specify all of the contingencies that may affect outcomes in given employment relationships. This creates a situation of "contractual incompleteness" in which exchange participants must instead agree on who will decide what to do, and under what limitations, when a contingency not covered by the contract is realized.

Alternatively, it may be difficult or impossible to specify contractually some important aspects of given employment relationships, simply because one or the other party to the exchange is either unable to observe whether given hypothetically acceptable terms have been fulfilled, or unable to provide verifiable evidence to external enforcement agencies of non-fulfillment of such terms. In this case, to the extent such aspects of the relationship are to be realized, incentives must be provided within the structure of the relationship rather than through externally applied penalties.

In the specific context of employment relationships, such incentive provisions may be necessitated by the difficulty of ascertaining workers' levels of labor intensity or effort. The need to provide effort incentives provides an appealing explanation for the empirical incidence of compensation schemes linking pay to some aspect of performance, including piece rates, commissions, bonus systems, promotion ladders, and the like. In all such cases, incentives for higher or more effective labor effort are provided through some combination of penalties for poor outcomes or rewards for evidence of superior performance. It also explains why employers invest resources in labor supervision and often base compensation on supervisory assessments of performance. The mainstream literature on the "principal-agent problem" (see Laffont and Martimort 2001 for an overview and synthesis) analyzes the structure of optimal incentive schemes under conditions of asymmetric information.

Although the presence of incentive problems introduces a source of inefficiency and may add new features to the structure of employment transactions, it does not have necessary implications for the mechanics of demand and supply in labor markets, so that expected compensation levels may adjust to clear these markets even if incentive constraints are binding and labor effort levels are inefficiently low. For example, a standard implication of second-best contracts derived in principal-agent models is that risk-averse agents receive their competitively determined utility levels even if they are induced to supply inefficiently low levels of effort due to the need to trade off effort incentives with insurance against production risks. The key to this result is that agents are made to bear payoffs below their reservation levels for "low" values of stochastic production outcomes, but are just sufficiently compensated by higher payments for "high" values to ensure that they attain their reservation payoffs in expected terms.

This story fundamentally changes, however, with attendant implications for the functioning of labor markets, if employers are somehow prevented from making workers bear below-reservation payoff levels for unfavorable production outcomes. In this case, if workers must thus receive at least their reservation payoffs for poor stochastic outcomes, then the only way for employers to provide incentives for higher labor intensity is to reward favorable outcomes with above-reservation level payoffs, implying that workers in this scenario must receive expected payoffs above their competitively-determined reservation utilities.

This outcome provides one theoretical basis for the extensively studied scenario of "efficiency wages" (see, for example, Weiss 1990 for models and references). There are other theoretical motivations for the payment of efficiency wages—for example, to reduce costly worker turnover, meet fairness norms or promote worker morale—but in all such scenarios, the presumption is that employers must pay supra-competitive wage levels to ensure the attainment of employment conditions that cannot be secured more directly through specification and external enforcement of appropriate contractual language. As we discuss below, this possibility has important qualitative implications for labor market outcomes.

2. Monopsony power

Our discussion of the implications of monopsony power in labor markets begins with the standard partial equilibrium textbook models, in which the presence of buyer-side market power is indicated in terms of the number of labor buyers in a given market. Consistently with the discussion in the previous section, the focus of the analysis then shifts to the role of matching frictions in an economy with a continuum of identical labor buyers.

2.1 *Pure monopsony*: *single-wage and wage discrimination cases*

To establish a frame of reference and some benchmark results, we start with the textbook case of a single firm purchasing labor in a market supplied by wage-taking sellers. Define L as the total quantity of labor demanded by the firm and denote the firm's revenue product function by $R = \theta f(L)$, with parameter $\theta > 0$ and f a twice continuously differentiable function such that $f'(L) \leq 0$. If the firm is a competitor in its output market, then θ represents its output price and f denotes its production function; otherwise, θ represents a shift parameter for the firm's revenue product function. Suppose that the firm faces a twice continuously differentiable market labor supply function $L = \lambda(w)$ with $\lambda'(w) > 0$ for all w and wage elasticity $e_{Lw} = \eta(w) > 0$. Let $\omega(L)$ denote the corresponding inverse market supply function, $\lambda^{-1}(L)$.

How would the monopsonist set wage and employment levels so as to maximize its profit? For the sake of comparison, note that if the firm's marginal revenue product function, $\theta f'(L)$, also corresponds to the market demand function for a perfectly competitive labor market, then the competitive equilibrium wage and employment levels would satisfy $\theta f'(L_c) = \omega(L_c) = w_c$. If the single labor buyer could exercise its monopsony power to the fullest extent consistent with the position of the market supply function, it would engage in perfect or *first-degree* wage discrimination and pay each unit of labor hired the corresponding reservation wage rate. In this case, the market supply function is also the firm's marginal factor cost curve, so its profit-maximizing level of employment would also be equal to L_c, but the firm would appropriate the entire area corresponding to producer and consumer surplus in the competitive equilibrium. Since the monopsony is able to appropriate the full surplus in this case, it has an incentive to maximize the surplus, so the outcome is efficient.

Suppose, however, that price discrimination were prohibited or the firm found it too costly to ascertain the reservation wage for each unit of labor it hired, and as a result the monopsonist sets a single wage rate. The monopsony's profit function is then given by

(1) $\pi(L) = \theta f(L) - \omega(L) \cdot L$.

The corresponding first-order condition for the profit-maximizing employment level L_m implies

(2) $MRP_L = \theta f'(L_m) = \omega(L_m) + \omega'(L_m) \cdot L_m = MC_L$,

as illustrated in Figure 1.

FIGURE 1 HERE

The firm's optimization condition can be manipulated to yield an expression for the *rate of monopsonistic exploitation*, *E*, understood as the proportional divergence of the monopsony wage from the marginal revenue productivity of labor:

(3) $E = (MRP_L - w_m) / w_m = 1/\eta(w_m)$,

indicating that the rate of monopsonistic exploitation is inversely related to the wage elasticity of supply evaluated at the profit-maximizing wage level. This is not precisely a statement of cause and effect unless wage elasticity is parametric or at least a monotonic function of demand parameters. For example, for the set of affine supply functions $L = a + bw$, the wage elasticity at any positive wage level is strictly decreasing in *a* and strictly increasing in *b*.

This result can be immediately extended to the case of *third-degree* wage discrimination, in which the monopsony is presumed able to identify separate segments of the labor market, distinguished by the elasticity of labor supply. The informational requirements for this form of wage discrimination are considerably less onerous than for first-degree discrimination, and to that extent it has greater presumptive empirical applicability. Note that this differentiation has no necessary implications for productivity, so the revenue productivity of labor remains solely a

function of total labor hired across all submarkets. The corresponding expression for the rate of monopsonistic exploitation in a representative submarket *s* is

(4) $$E^s = (MRP_L - w_s) / w_s = 1/\eta^s(w_s)$$

where $\eta^s(w_s)$ denotes the elasticity of wage supply in submarket *s* evaluated at its equilibrium wage rate. Subject to the caveat about causality previously noted, this extension indicates that that the monopsonist exploits labor more intensively the more inelastic is the supply of labor. Intuitively, the monopsonist is able to do this because greater inelasticity suggests that workers in that submarket have less viable alternatives at the margin to working for the firm.

As a final extension of the pure monopsony scenario, we anticipate the subsequent discussion of labor flows by adding a time dimension to the monopsonist's optimization problem. Denoting time periods by the subscript *t*, we write the quantity supplied of labor available to a monopsonist in period *t* as

(5) $$L_t = (1 - \sigma(w))L_{t-1} + \rho(w),$$

where $\sigma(w) \leq 1$ represents the rate at which incumbent workers leave the firm in any period, understood to be a decreasing differentiable function of the wage rate, and $\rho(w)$ is the rate at which the firm can recruit new workers, a differentiably increasing function of the wage rate.

In a steady state, flows into and out of the firm must balance to ensure $L_t = L_{t-1}$ for all *t*, implying that the steady-state labor supply to the firm is given by

(6) $$L(w) = \rho(w) / \sigma(w),$$

with corresponding steady-state labor supply elasticity given by

(7) $$\varepsilon_{Lw} = \varepsilon_{\rho w} - \varepsilon_{\sigma w},$$

the difference between the recruitment and separation elasticities. This can also be thought of as the *long-run* elasticity of labor supply to the firm, insofar as the desired level of steady-state employment is entirely a choice variable for the labor buyer.

In contrast, L_{t-1} is given in the short run, and so, in light of (5), the corresponding short-run supply elasticity is expressed as

(8) $$\varepsilon^{s}_{Lw} = -\varepsilon_{\sigma w}\sigma(w_t)\cdot(L_{t-1}/L_t) + \varepsilon_{\rho w}\cdot(\rho(w_t)/L_t).$$

Noting again that $L_t = L_{t-1}$ in the steady state and combining (6), (7), and (8) yields

(9) $$\varepsilon^{s}_{Lw} = \sigma(w_t)\varepsilon_{Lw},$$

indicating that the short-run wage elasticity of supply is generally less elastic than its long-run counterpart.

How do the short- and long-run supply elasticities figure into the monopsonist's intertemporal optimization problem? Given the firm's time discount factor $\delta \leq 1$, it can be shown that in the steady state the monopsonist maximizes its present discounted value of the profit stream by setting the wage rate so that

(10) $$E = (MRP_L - w)/\ w = [(1-\delta)/\varepsilon^{s}_{Lw}] + [\delta/\varepsilon_{Lw}],$$

that is, the rate of exploitation is equal to a weighted average of the short- and long-run inverse wage elasticities of supply, where the weights are determined by the firm's discount factor (Boal and Ransom 1997, p. 90).

Taking stock of the argument to this point, the key implications of the pure monopsony scenario are as follows: (i) The primary effect of the pure exercise of monopsony power is redistribution rather than inefficiency, insofar as a first-degree wage discriminating monopsonist achieves the efficient level of employment by appropriating all of the available market surplus. Thus, to derive inefficient outcomes from the expression of monopsony power, it is necessary to

adduce some additional market frictions, such as negotiation costs which make this degree of wage discrimination unappealing. (ii) A profit-seeking, single-wage setting monopsonist will in general set wage and employment levels inefficiently low, the proportional divergence of wage from the revenue productivity of the last unit of labor hired being inversely related to the wage elasticity of supply. In such cases, there is a potential role for government intervention to boost the efficiency of market outcomes. For example, by imposing a minimum wage above the monopsony equilibrium level, the government forces up the firm's marginal factor cost curve. If the minimum wage is set at the perfectly competitive equilibrium level, the monopsonist's marginal cost is horizontal at that point, and the firm maximizes profit by setting employment and the wage rate at their competitive equilibrium levels. Another possibility, based on theoretical grounds discussed later in the chapter, is for the government to offset monopsony power by supporting collective bargaining by labor suppliers. (iii) Monopsonistic wage discrimination violates the competitive hypotheses of LOW and CDW, insofar as wage differences arise which do not correspond to underlying variations in worker productivity or working conditions. (iv) The pure monopsony model provides no basis for inferring the existence of unemployment, voluntary or otherwise. The marginal worker hired receives her reservation wage, and no unhired worker is willing to work at the going wage rate.

2.2 *Oligopsony*: *partial and general equilibrium views*

Practically speaking, cases of pure monopsony are rare, so that to the extent it exists, monopsony is best thought of as a matter of degree. In other words, *oligopsony* is presumptively the general scenario for buyer-side wage-setting power in labor markets. This leaves open the question of how the manifestation of oligopsony power is to be understood.

Here, we present two approaches to this question. In the first, traditional approach, the degree of oligopsony is represented parametrically by the number of labor buyers in a market with a given labor supply curve. We then present a more general framework due to Burdett and Mortensen (1998) in which the presence of oligopsony is determined by matching frictions in a flow economy, rather than by the assumed paucity of firms in a given labor market. As a consequence, the degree of wage-setting power expressed in equilibrium is a function of limits to worker mobility implied by giving restrictions in the job matching process rather than the number of firms, which would be consistent with competitive equilibrium in the absence of such frictions.

We can build the former approach as a continuation of the pure monopsony case considered above by assuming there are *n* identical labor buyers indexed by $i = 1, 2, ..., n$, each with revenue product functions $R^i = \theta f(L_i)$, with the same properties as before. In parallel fashion, we assume that the inverse market labor supply function is given by $w = \omega(L)$, where $L = \sum_{i=1}^{n} L_i$. If the single market wage rate is determined endogenously by the employment choices of the labor buyers in the market, the profit function of a representative firm is given by

$$\pi(L_i) = \theta f(L_i) - \omega(L_i + \sum_{j \neq i} L_j) \cdot L_i \qquad (11)$$

Thus we have specified a non-cooperative game with *n* identical players, each of whom chooses an employment level L_i as a strategy, and has the payoff function (11) defined as a function of all players' strategies. Along with the assumptions adopted so far, a unique Nash equilibrium in employment strategies exists for this game if each player's payoff function is strictly concave in her own employment choice and furthermore the players' employment choices are strategic substitutes for each other, in the specific sense that player increasing player *j*'s employment level

increases the marginal factor cost of labor to player $i, i \neq j$ and thus reduces *i*'s *marginal* payoff—that is,

$$(12) \qquad \partial^2 \pi(L_i) / \partial L_i \partial L_j = -[\omega'(L_i + \sum_{j \neq i} L_j) + \omega''(L_i + \sum_{j \neq i} L_j) \cdot L_i] < 0 .$$

Given the strategies of the other players, each firm's optimizing employment choice L_i^* satisfies

$$(13) \qquad MRP_L^i = \theta f'(L_i^*) = \omega(L_i^* + \sum_{j \neq i} L_j) + \omega'(L_i^* + \sum_{j \neq i} L_j) \cdot L_i^* = MC_L^i$$

Given the symmetry of the players' payoff functions, the unique Nash equilibrium is such that each firm hires the same amount of labor, and thus the equilibrium rate of exploitation for the market is given by

$$(14) \qquad E(n) = (MRP_L - w(n)) / w(n) = 1 / (\eta(w(n)) \cdot n)$$

where $w(n)$ denotes the Nash equilibrium wage level as a function of the number of labor buyers in the market. Note that for any non-zero wage elasticity of labor supply, the rate of exploitation is strictly decreasing in the number of labor buyers, and approaches zero in the limit as the number of buyers approaches infinity. Thus, so long as the market supply curve is upward-sloping, the number of firms provides an index of the degree of labor market competitiveness.

Readers familiar with industrial economics will recognize this result as parallel to that for the Cournot model of oligopoly in product markets, and so can anticipate that this result is not robust to alternative specifications of oligopsonistic competition. In the Bertrand oligopsony scenario, for example, strategic wage choices between homogeneous firms with constant marginal revenue product functions yield competitive equilibrium wage and employment levels with two or more labor buyers in the markets. The Bertrand model thus underscores the point that worker mobility among buyers, rather than the number of buyers per se, is the key limiting factor to monopsony power when buyers choose the terms of labor exchange strategically.

This insight lies at the heart of the alternative approach to modeling oligopsony due to Burdett and Mortensen (1998), which derives the degree of monopsony power from the degree of effective mobility in an economy characterized by labor flows into and out of the labor market. Here we present the key structural elements of their approach along with the main results, referring the reader to their article, or to simplified versions of their argument in Boal and Ransom (1997) and Manning (2003, Chapter 2).

The Burdett-Mortensen model of labor flows features a large number of identical workers who are either employed or unemployed in a given period; a large number of identical firms each with constant returns to scale production using only labor; a fixed flow of job offers per period to workers, understood as random draws from a wage distribution to be determined within the model; an exogenously given rate of destruction of existing employment matches, which creates newly unemployed workers; and a fixed payoff per period to unemployed workers, determined say by the level of unemployment benefits.

The assumed behavior of workers and firms is very simple. Unemployed workers accept any wage offer at least as great as their reservation wage, which is just equal to their payoff to being unemployed. Employed workers accept any wage offer higher than their current wage. Given worker behavior, the fixed job offer and job destruction flows, and the overall wage distribution, each firm chooses its wage rate to maximize its long-run profit based on its steady-state employment level, determined as in equation (6) above, given the wage distribution $F(w)$. Equilibrium for the economy as a whole entails that all firms receive identical profit levels.

Since firms are identical and choose wages strategically, competition among firms would drive the equilibrium wage to the competitive level for all workers if there were no mobility restrictions. But each period some workers find themselves unemployed by external forces, and

while all workers receive wage offers each period, not all employed workers necessarily receive offers that are at least as good as their present wages. Thus, depending on the form of the equilibrium wage distribution, some employers may be able to offer lower than competitive wages without losing their entire labor force.

The analytical task on the basis of this model is ultimately to determine the steady state cumulative wage distribution $F(w)$, which will also allow calculation of the expected equilibrium wage rate. Burdett and Mortensen's most startling result is that, even though all workers and all firms are assumed to be respectively identical, equilibrium is such that there is a continuous distribution of wages ranging between the payoff received by unemployed workers and an upper bound that is below workers' marginal revenue product and strictly decreasing in the unemployment rate determined by the exogenous job destruction and job offer flows. Mobility restrictions explain the latter result: to the extent that workers become unemployed but can't immediately garner competitive wage offers, all firms will enjoy monopsony power.

The more startling result, given that firms and workers are assumed to be respectively identical, is that the equilibrium wage distribution is nondegenerate. This result flows from the assumed mobility restrictions plus the presence of constant returns in production. To see this, suppose instead that equilibrium were such that a positive measure of employers offered the same wage. Then one of those employers could, by offering an infinitesimally higher wage rate, draw a positive measure of new recruits, thus significantly increasing output while lowering profit per unit of output only slightly. But then the original wage distribution could not be consistent with equilibrium, since firms would have an incentive to alter it.

Burdett and Mortensen's "general equilibrium" approach based on mobility restrictions yields other important insights about the context in which monopsony power arises in labor

markets. First, there is equilibrium unemployment in their model, although this is driven mechanically by the assumption of exogenously given job destruction and job offer rates. Second, and more significantly, they show that the expected wage level in the economy will be a weighted average of the minimum possible wage paid to workers, established by the payoff to unemployment, and workers' common marginal revenue product, with the weight on the former given by the employment rate. This result yields as a corollary the "wage curve" hypothesis of an inverse relationship between real wage levels and unemployment rates in given labor markets (see Blanchflower and Oswald 1994). In the context of the present discussion of imperfect competition, this result is significant because it is at odds with the CWD hypothesis as it pertains to the existence of unemployment in given labor markets. As Blanchflower and Oswald (1994, pp. 37-38) discuss, if behavior were otherwise competitive, workers faced with a higher risk of being unemployed in a given market would need to be compensated with a higher wage rate, contrary to the prediction of Burdett-Mortensen model.

The model also provides an explanation for the longstanding empirical puzzle in labor economics concerning the wage premium paid by large employers, holding workplace and worker characteristics constant (Troske 1999). In this context, the wage premium is a necessary byproduct of equilibrium wage dispersion and constant returns production, given the equalization of profits in equilibrium: in order to compensate for making lower profit per worker in equilibrium, high-wage firms hire more workers and produce correspondingly higher levels of output. Similarly, the model provides a theoretical basis for the empirical regularity that workers with more experience or tenure receive higher wages (Altonji and Williams 2005), but with a twist: in contrast to the human capital explanation that such workers earn higher wages due to

higher productivity, in the present context higher-wage workers tend to have higher tenure simply because they are thus less likely to get superior wage offers from other firms.

As Manning (2003) points out, the model also provides a basis for assessing the level of effective monopsony power in real-world labor markets characterized by worker flows across firms and between non-employment and employment. As suggested by the expression for the expected value of equilibrium wages, employers compete head-to-head in setting wages to the extent that recruited workers are drawn from other firms rather than from the pool of non-employed workers. As this the proportion of new hires coming from other firms approaches one, the expected wage rate collapses to the competitive level, and as it approaches zero, the expected wage rate falls to the lowest possible level. Consequently, a statistic measuring the proportion of new hires coming from non-employment provides a measure of the extent of wage-setting power enjoyed by firms. Manning argues that data from the US and UK show not only that a substantial proportion of new hires come from non-employment, but that variations in this proportion over the business cycle and across demographic groups are consistent with the wage variations predicted by the model (Manning 2003, pp. 44-49). This type of indirect evidence of monopsony power is valuable insofar as direct evidence of monopsony power's effect on wages and employment is suggestive for specific labor markets but mixed overall (Boal and Ransom 1997).

What are the normative implications of the Burdett-Mortensen framework? Although the equilibrium is clearly Pareto-inferior relative to competitive equilibrium, given the existence of unemployment in the steady state, this is not really a relevant comparison—if the economy is beset by matching frictions as the model presupposes, then the competitive equilibrium is not attainable in any case. The appropriate question is whether the equilibrium constitutes a second-best outcome in the presence of such frictions, or whether there is scope for Pareto-improving

interventions such as minimum wages. This is not necessarily the case, since simply raising the lower bound of the equilibrium wage distribution does not increase the rate at which workers are matched with firms. As in the partial equilibrium scenario of oligopsony in a market with a rising labor supply curve, the presence of worker heterogeneity in reservation wage rates implies that expected wages and the equilibrium number of hires are too low, so that an effective minimum wage would improve efficiency in this case. But this conclusion is not robust to other extensions of the model including free entry by firms. The appropriate role of government policy vis-à-vis labor markets is thus ultimately an empirical rather than a purely theoretical issue.

3. Efficiency wages

3.1 *A basic model with job separations and firing for cause*

Consider again a labor market in which a representative firm has the revenue product function described at the beginning of section 2.1, but now allow that there are sufficient firms in the market (and sufficiently low matching frictions) to assure that the buyer side of the market is otherwise competitive. In addition, and critically, suppose that effective labor input L is given by the product el, where l is the number of worker hours as before, but e represents labor intensity or effort. For the sake of simplicity, let e take two values, $e_L = 0$ and $e_H = 1$, so that the higher effort level is required in labor market equilibrium.

Under the assumptions of the efficiency wage framework, as discussed earlier, employers can neither automatically see worker effort nor infer individual effort levels from output. Consequently, firms can only induce workers to choose higher effort by monitoring their work activity and threatening to fire workers who are caught working at less than the desired intensity

level. In this scenario, workers' effort incentives stem from the desire to avoid the lower payoff stream associated with being fired.

Suppose that all workers are identical, and let a representative worker's immediate payoff from receiving wage *w* and performing effort *e* be given by the simple function $U = w - e$. Suppose that each worker discounts future payoffs by the fractional discount factor δ, and let V_E and V_U respectively denote the present value of the utility stream from being currently employed and the present value of the utility stream from being newly unemployed, both measured at the beginning of a period. Let the fraction *p* denote the proportion of the incumbent labor force that whose jobs are eliminated, independently of effort choices, and assume initially that this is exogenously given. Then the present value of the representative worker's utility stream from receiving *w* and performing $e = 1$, denoted V_1, is given by

$$V_1 = w - 1 + \delta[p V_U + (1 - p)V_E] \tag{15}$$

Suppose further that a given worker who elects to perform low effort $e = 0$ is dismissed with fractional probability $q = p + s$, where *s* is understood as the probability that the worker is detected performing low effort and is then fired for cause. Then the present value of the worker's utility stream from receiving *w* but performing $e = 0$, denoted V_0, is expressed by

$$V_0 = w + \delta[q V_U + (1 - q)V_E] \tag{16}$$

In order to induce each worker to perform high effort, firms must choose their payment streams to satisfy the *incentive compatibility* constraint $V_1 \geq V_0$, which implies, after substituting expressions (15) and (16) into the constraint, $V_E - V_U \geq MC_e / \delta s$, where MC_e (=1 in this case) denotes the worker's marginal disutility of performing higher effort. Profit-seeking firms have no

incentive to offer workers more than the minimum payoff to induce high effort, so in equilibrium the constraint is satisfied with strict equality, yielding

(17) $V_E = V_U + MC_e / \delta s$

Now let *b* denote a worker's immediate payoff to being newly unemployed; depending on the context, this might represent the value of leisure augmented by any unemployment compensation, or the worker's payoff from being employed in some other sector of the economy. Assume that unemployed workers are reemployed with some matching probability μ. Then in a steady state a worker's present value of the utility stream from being newly unemployed is

(18) $V_U = b + \delta[\mu V_E + (1-\mu)V_U]$.

Finally, given satisfaction of the incentive compatibility constraint, workers always perform high effort in equilibrium, so that (15) becomes

(19) $V_E = w - 1 + \delta[p V_U + (1-p)V_E]$

Solving the system of equations (17)-(19) for the steady-state "efficiency" wage rate yields $w = b + e_H + \{[(1-\delta) + \delta(p+\mu)]MC_e\} / \delta s$, where $e_H = 1$.

To appreciate fully the implications of this derivation for labor market outcomes, consider one additional nuance which incorporates market flows more fully in the analysis. Suppose that there is some probability *m* that a given unemployed worker is successfully matched with a new employer, and let *u* be the unemployment rate. Then $\mu = m / u$, and the equilibrium wage rate satisfies

(20) $w_e = b + e_H + \{[(1-\delta) + \delta(p + (m/u))]MC_e\} / \delta s$

This outcome is illustrated in Figure 2. For comparison, the competitive equilibrium wage rate in this scenario would be $w_c = b + e_H$, as this would be the payment just sufficient to cover an employed worker's outside payoff *b* in addition to providing a compensating differential for the

performance of high labor effort. The wage required for satisfaction of the incentive compatibility constraint, in contrast, exceeds the competitive wage by a factor that is increasing in the probabilities of being subsequently matched if unemployed and of being dismissed other than for cause, and decreasing in the unemployment rate and the probability of being fired for cause if caught performing low effort.

[Figure 2 Here]

If we assume for the sake of illustration that there are *N* workers in the market and there would be exactly full employment in competitive equilibrium, then the unemployment rate, given that L_e workers are hired in the efficiency wage equilibrium, is given by $u = (N - L_e)/N$. This determines the incentive compatibility (IC) locus, defined as the set of wage and employment levels that satisfy (20) given the expression for the unemployment rate.
As Figure 2 indicates, this equilibrium departs from the competitive outcome in a number of significant ways. First, most immediately and in conflict with NUV, there is involuntary unemployment in equilibrium, in the sense that there are unemployed workers who would be willing to work at or below the equilibrium wage, but cannot get employment because there is no unsatisfied quantity demanded at that wage, and they can't credibly promise to exert the high effort level at any lower wage. Second, there is a wage curve, as the equilibrium wage rate is declining in the unemployment rate.

Third, there are potential conduits for the operation of worker- and firm-specific effects on the equilibrium wage that are inconsistent with LOW and CWD. For example, differences in the rate at which workers discount future payoffs or are matched into new jobs will affect the equilibrium wage, and on the firm side, variations in layoff or firing probabilities will induce differences in labor market outcomes for workers with identical productivity characteristics.

As discussed by Weiss (1990), there are other versions of the efficiency wage model in addition to the account based on effort incentives summarized here. In these other models, firms' motivation to pay supra-competitive wages derives from the desire to reduce turnover, promote a queue of relatively high-quality workers when these qualities can be only imperfectly assessed at the time of hiring, or to increase and maintain worker morale. The common denominator in all these accounts is that the efficiency wage serves as a strategic response to the difficulty of securing desirable traits or behaviors through the use of targeted penalties enforced by contract.

3.2 *Do (any) firms pay efficiency wages*?

The efficiency wage hypothesis has received considerable attention in the economic literature, on both theoretical and empirical grounds. On the theoretical level, there has been much debate in the literature concerning the cogency of the premise that firms must pay supra-competitive wages to address imperfect contracting conditions, particularly with respect to the effort incentive problem. For example, some have argued that employers can require employee bonding against unfavorable outcomes, or charge newly hired workers employment fees or bonds equal to their expected employment rents, or place workers in entry-level jobs which pay below-competitive wage rates while promising eventual promotion to jobs with supra-competitive compensation. To date, empirical evidence on the incidence of efficiency wages, like that on the manifestation of monopsony power, is suggestive but mixed (see Weiss 1990, pp. 6-13, Manning 2003, pp. 125-128, and St. Paul 1996, p. 95 for references and summaries of the theoretical and empirical evidence.)

3.3 *Dual labor markets*, *internal labor markets*, *and wage rigidity*

Efficiency wage theory has also been used to provide microfoundations for the hypothesis of *dual* or *segmented* labor markets (Bulow and Summers 1986, St. Paul 1996). The duality hypothesis posits a discontinuity in the types of jobs available to workers, such that *primary sector* jobs are characterized by relatively high wages, job security and tenure, and good nonpecuniary benefits and working conditions, while jobs in the *secondary* labor market pay relatively low wages, offer few or no benefits, and have high turnover. Crucially, these differences are understood to be experienced by workers who, based on available empirical measures, would be treated equally under competitive conditions.

Labor market duality is thus essentially at odds with the LOW and CWD hypotheses, and any candidate analytical framework for the hypothesis would need to explain why worker mobility across jobs doesn't eliminate such disparities. As noted above, efficiency wage theory provides an explanation for such discontinuities in equilibrium by adding the postulate that primary-sector jobs are typically characterized by the sort of contractual failures that require firms to offer supra-competitive terms to secure profit-maximizing worker behaviors or types. Workers in the secondary sector with equal qualifications would be willing to work in the primary sector jobs for less favorable terms, but because of the contractual imperfections cannot credibly promise to provide the desired qualifications or behavior.

Another feature associated with primary-sector jobs is apparent relative immunity of equilibrium wage and hour levels to variations in market conditions. This asserted invariance is part of a more comprehensive hypothesis concerning the existence of *internal labor markets* (Doeringer and Piore 1971, Osterman 1984) in which job descriptions, promotion ladders and wage structures are set administratively by firms, and then do not respond at the margin to external shocks to marginal productivity, as would be predicted by competitive theory. This

phenomenon suggests an additional, qualitative distinction in the logic of wage determination in the primary and secondary segments of the labor market.

Saint-Paul (1996) constructs an explanation for this phenomenon by adding a dynamically stochastic element to the standard efficiency wage model based on effort incentives. To see how his account works, suppose that the shift parameter θ in the firm's revenue product function is understood to vary randomly across time periods, so that each firm finds it necessary to reconsider its labor market choices in each period as new economic conditions arise. To capture this, add a time subscript *t* to the shift parameter and the firm's wage and employment levels, and suppose that successive values of θ_t are drawn from identical independent distributions such that the firm knows the value of the shift parameter one period in advance. The firm thus chooses its period-*t* labor and wage levels to maximize its profit in that period plus the expected present value of the future stochastic stream of profits, given that the firm will make similar dynamic responses to subsequent variations in the shift parameter. The question thus becomes how, in a given period *t*, the firm will choose to alter its current-period choices w_t and L_t in response to new market conditions in period *t*+1.

Given the effort incentive condition, the expression for the efficiency wage in period *t* becomes

$$w_{t+1} = b_{t+1} + e_H + \{[(1-\delta) + \delta(p_{t+1} + (m_{t+1} / u_{t+1}))]MC_e\} / \delta s, \tag{21}$$

where the time subscripts on *b*, *p*, *m* and *u* indicate that these parameters may be influenced by random shocks to revenue productivity in the primary sector. Now consider the behavior of the job elimination probability, p_{t+1}, and its implications for the firm's wage and employment decisions in period. In the dynamic context, p_{t+1} must be considered at least in part endogenously determined, since in each period the firm must decide how to alter its incumbent labor force in

response to the external shock. Note that there is an asymmetry at the firm's current level of employment, L_t, since decreasing the incumbent labor force requires an increase in p_{t+1} but increasing the labor force does not. As can be seen from expression (21), an increase in p_{t+1} implies an increase in the wage that must be paid in order to satisfy the incentive compatibility constraint, other things equal. Consequently the firm's marginal cost of increasing its existing labor force is higher than its *reduction* in cost from downsizing the labor force.

The result, as shown in Figure 3, is a "kink" in the firm's labor cost function at the period-*t* level of employment, with a corresponding discontinuity in the associated marginal cost curve. This discontinuity yields two implications for the firm's choices of wages and employment in response to external shocks. First, small downward shocks, which cause variations within the gap between the low and high levels of marginal cost, induce no changes in the firm's existing wage and employment levels, since no other combination of these levels leads to higher profit. This establishes a microeconomic basis for wage and employment rigidity, despite changing external market conditions.

[Figure 3 here]

Second, relatively large downward economic shocks, sufficient to push the firm's marginal revenue productivity to the lower level of marginal cost, will have permanent effects in the sense that a new marginal cost discontinuity will be established at the lower employment level. This introduces an element of *hysteresis* into the labor market, in which purely temporary shocks can have long-term consequences for employment.

4. Bargaining[3]

Monopsony and efficiency wage models are consistent with the observation that employers typically set wage levels, at least in the US, and particularly at the point of initial hires. However, there are a number of cases in which labor suppliers presumably also have some voice in the determination of wages and working conditions. The most obvious case is when workers are represented in collective bargaining arrangements that are guaranteed by law, as with trade unionization. However, intuition and observation suggest that wage-setting power is also enjoyed by workers with unique talents and abilities (as in labor markets for professional sports teams). In addition, just as worker mobility costs create monopsony power for labor buyers, the cost of replacing incumbent workers give the latter countervailing "insider power" relative to workers being hired directly from the labor market. In such cases, it may be more appropriate to represent the process of wage determination within a bargaining framework. In this section, we introduce such a framework and discuss some distinctive implications of bargaining models for labor market outcomes.

4.1 *Strategic wage bargaining*: *a basic model*

Perhaps the key problem to be addressed in the economic analysis of bargaining is how to capture in a tractable and coherent choice-theoretic framework the general intuition that outcomes depend on the relative bargaining power of the negotiators, and further how to do so in a manner that yields sharp testable hypotheses. This problem is illustrated by the textbook case of bilateral monopoly, which involves combining a monopoly union model with the pure monopsony model illustrated in Figure 1. As is well known, the model does not yield determinate predictions about wage or employment levels.

A traditional response to this problem in the economics literature involved the use of the *Nash bargaining solution* of cooperative game theory, which Nash derived from a set of axioms understood to characterize the outcomes that might be expected from the interactions of rational, self-seeking players who can make binding commitments to any agreements reached through negotiation. This solution, which involves the distribution that maximizes the product of the players' net payoffs, has the property that each player's bargaining payoff is strictly increasing in the payoff he or she can expect to receive in case the bargaining relationship breaks down, and strictly decreasing in the corresponding outside payoff of his or her opponent.

The problem with this solution, as with other equilibrium concepts derived using the axiomatic approach, is that there is no way within the framework to assess the relevance of the particular axioms adopted, and similarly, no clear way to assess the strategic significance of the parameters of the model. For example, in the Nash bargaining solution, it is unclear how to interpret the conditions under which players receive their "outside" payoffs—that is, whether these are elected voluntarily by the players under some circumstances, or imposed exogenously. Similarly, while the basic Nash solution can be modified to allow for differential bargaining power of the players, there is no way within the analytical framework to identify the strategic foundation of such differentials. To address this problem, Nash proposed an analytical response, now referred to as the "Nash program," in which given cooperative-game solutions are validated by demonstrating their emergence as equilibria for appropriately specified noncooperative or strategic games.

An important step forward in this program was provided by Rubinstein (1982), who established a framework for deriving unique and intuitively appealing bargaining outcomes as subgame-perfect equilibria of strategic games in which players alternate making and responding

to offers over time until an agreement is reached. In this framework, the relative bargaining power of the players is determined by the form and respective levels of the costs they incur in engaging in successive rounds of bargaining.

Rubinstein's approach makes crucial use of the recursive structure of these games. To see how this works, suppose that two parties, a firm F and a worker W, bargain over the gains *S* to be generated by a prospective employment relationship between them, with the understanding that they will bind themselves to any agreement reached through negotiation. Players are risk-neutral and care only about their own bargaining shares. Let *x* denote the share of the surplus received by W in the event of agreement, so that $S - x$ is F's share. Suppose further that bargaining proceeds over discrete time periods $t = 0,1,2,....$, with W making the initial offer in period zero, which F either accepts or rejects. If the latter, F makes a counter-offer in period 1, W in turn either accepts or rejects. Bargaining proceeds in this manner, with players alternating in making and responding to offers, until agreement is reached. Assume finally that each player discounts future payoffs by fractional discount factor $\delta_i, i = F,W$, so that a bargaining payoff z expected by player *i* in the next bargaining round is evaluated at $\delta_i z$ in the current round.

A segment of the alternating-offer game just described is illustrated in Figure 4. This illustration makes clear the recursive structure of the game, such that the bargaining game beginning at any even-numbered period exhibits the same ensuing sequence of moves, with player W making the initial offer to which F responds, etc. Given this periodic recursion, Rubinstein demonstrates that unique equilibrium payoffs can be derived via analysis of the players' reaction functions for any given two-period sequence beginning with an offer by W.

[Figure 4 here]

Define F's bargaining reaction function as the value *y* of the offer made by F that is minimally acceptable to W, given that W expects a payoff of *x* in the following bargaining period. By this definition, F's bargaining reaction function is given by the equation $y = \delta_W x$, reflecting W's time cost of rejecting F's current offer and receiving the expected payoff of *x* in the next period. Similarly, remembering that F receives the net surplus $S - x$ if W receives *x*, define W's bargaining reaction function as specifying the value of *x* proposed by W that is minimally acceptable to F, given that F expects to receive a payoff of $S - y$ in the following period. This condition is met for the game specified above by the expression $S - x = \delta_F \cdot (S - y)$.

The equilibrium bargaining payoff to W is derived by the value of *x* that solves this equation system, denoted x^*, with the equilibrium payoff to F given by $S - x^*$. As Rubinstein shows, these payoffs are uniquely implemented by a pair of strategies in which W initially proposes to receive x^* and F immediately accepts.

The solution to the reaction functions specified above entails the wage

(22) $$w = x^* = (1 - \delta_F)S / (1 - \delta_F \delta_W),$$

with F receiving the remainder $S - x^* = \delta_F (1 - \delta_W)S / (1 - \delta_F \delta_W)$. This solution is illustrated in Figure 5. It can readily be checked that each player's payoff is increasing in his or her discount factor, and decreasing in his or her opponent's, which accords with the intuition that more patient players do better in bargaining if bargaining takes time. In the case that $\delta_i = \delta \ \forall i$, the respective equilibrium payoffs are $\{S / (1 + \delta), \delta S / (1 + \delta)\}$, indicating a first-mover advantage that is decreasing in δ, with each player receiving half of the surplus in the limit as δ approaches 1. Similarly, the first-mover advantage is eliminated by replacing the arbitrary assumption that a

given player always goes first with the assumption that each player has an equal chance of making the initial offer.

[Figure 5 here]

Rubinstein also studies a version of the model with constant per-period bargaining costs c_i, which yields the sharply discontinuous equilibrium result that the player with the lower bargaining cost receives the entire surplus if making the first offer, and the entire surplus net of the other player's one-period bargaining cost if not. This discontinuity disappears if fixed and discounting costs are combined in the same model, though the unique equilibrium payoffs retain the implication that each player's payoffs are declining in her bargaining costs and increasing in her opponent's.

4.2 *The role of outside payoffs*: *exit vs. exogenous termination*

The Rubinstein framework has also been used to elucidate the meaning and equilibrium impact of "outside" payoffs on bargaining outcomes. One possible interpretation is that players realize their outside payoffs when the bargaining relationship is terminated as the result of an exogenous shock, rather than being voluntarily selected by one or the other bargainer. To analyze this scenario, let the fraction β denote the exogenous probability that the bargaining relationship breaks down if agreement is not reached in a given period, and let the players' respective outside payoffs be given by $a_i, i = F, W$. To ensure the economic viability of the bargaining relationship, assume that $S \geq (a_F + a_W)$. Then if we assume for notational simplicity that players have identical discount factors, their respective bargaining reaction functions take the form $y = \beta a_W + (1-\beta)\delta x$ and $S - x = \beta a_F + (1-\beta)\delta(S - y)$, which yield in turn the equilibrium bargaining payoff to W is $x^* = \{[1-\delta(1-\beta)]S + \beta[\delta(1-\beta)a_W - a_F]\} / [1-(\delta(1-\beta)^2]$. Taking the

limit of this expression first as δ approaches 1 and as β approaches zero, the worker's equilibrium wage becomes

(23) $$x^* = (S + a_W - a_F)/2,$$

with F receiving $S - x^* = (S + a_F - a_W)/2$. In this equilibrium, each player's bargaining share is strictly increasing in his or her own outside payoff, and strictly decreasing in the opponent's outside payoff. This is also the Nash bargaining solution for the utility possibilities frontier determined by S and the "breakdown" payoffs (a_F, a_W). Thus, strategic bargaining analysis indicates that the Nash bargaining solution emerges from a bargaining situation in which players discount the future very slightly (say, because bargaining rounds are very short) and there is a vanishingly small probability that the bargaining relationship will terminate in any period that agreement is not yet reached.

Suppose instead that the outside payoffs are realized only if one or the other parties voluntarily chooses to exit the relationship rather than accepting an existing offer. A player can credibly threaten to exit rather than accept a standing offer only his or her outside payoff is at least as high as the expected payoff to continued bargaining. Thus, if players discount future payoffs at the same rate, the bargaining reaction functions for this case are kinked, taking the respective forms $y = \max\{a_W, \delta x\}$ and $S - x = \max\{a_F, \delta(S - y)\}$. Due to these nonlinearities and the viability condition, there are three distinct equilibrium cases, depending on the values of the outside payoffs. Again taking the limit as the common discount factor approaches one, these cases are summarized as follows:

$$(24) \qquad x^* = \begin{cases} S/2 & if\ S/2 \geq \max\{a_F, a_W\} \\ a_W & if\ a_W > S/2 \\ S - a_F & if\ a_F > S/2 \end{cases},$$

with the firm receiving $S - x^*$ in each case. It remains the case in this scenario that bargaining concludes immediately in equilibrium, with the player making the initial offer proposing the equilibrium partition of the surplus, and the other player immediately accepting, so the exit options are never exercised.

Thus, in contrast to the outcomes predicted by the Nash bargaining solution, outside payoffs only affect bargaining outcomes if one or the other payoff is sufficiently high relative to the value-added generated by the bargaining relationship (they can't both be relatively high, due to the viability condition), in which case the party enjoying the relatively high outside payoff just receives that value in equilibrium, with the other player receiving the remainder. For example, according to this outcome, workers would enjoy asymmetric bargaining power in periods of low unemployment and high vacancies, while firms would enjoy this power at the other extreme of the business cycle when unemployment is high and vacancies low. In the intermediate range, however, bargaining payoffs would be determined solely by the parties' bargaining costs.

Both experimental economic and econometric evidence have been brought to bear on the question of whether individuals (and firms) behave according to the Rubinstein model. Binmore, Shaked and Sutton (1989) describe one such experiment that they conducted, which showed that outside options were important for determining splits, but the prediction that players would never play beyond a first offer was not supported. Scaramozzino (1991), using data from a panel of UK manufacturing firms sorted by bargaining regimes, did found that the structure of wage determination varied depending on the type of outside option regime, and that observed wages

were only affected by average industry wages in cases where an outside option was available (i.e., where industry wages affected employers' alternative profit opportunities and employees' alternative job prospects).

4.3. *Employee replacement costs and insider bargaining power*

As illustrated in the previous subsection, a unique implication of the bargaining approach to wage setting is that the outside payoffs of both employer and employees play a role in the determination of bargaining outcomes. In particular, and in keeping with competitive theory, both the Nash bargaining solution and the Rubinstein equilibrium with exit options predict for the 2-player case that the employer can capture virtually the entire net surplus if the firm can replace the incumbent worker at no cost, so that the employer's outside payoff is equal to the gross surplus minus the outside payoff to labor suppliers. Conversely, the incumbent worker gains bargaining power and a share of the net surplus to the extent that replacement is costly to the employer, and thus have an "insider" advantage relative to workers not currently in any firm's labor force (Lindbeck and Snower, 1989). Such replacement costs could derive from the loss of acquired firm-specific human capital or the difficulties of search, for example.

This story gains additional strategic dimensions once the framework is expanded to incorporate the typical case that firms typically employ multiple workers (implying a "monopoly-oligopsony" relationship rather than pure bilateral monopoly), so that there is an asymmetry in the respective exit options faced by employer and employees: while for each worker, "exit" entails a match with an entirely new employer, absent collective bargaining arrangements an employer can "exit" the bargaining relationship with a given employee simply

by making do with one less incumbent worker. This asymmetry has potential implications for both the firm's choice of production technique and the economic impact of collective bargaining.

These points are illustrated with a simple model of intrafirm bargaining with exit options due to Skillman and Ryder (1993).[4] In this model, production requires two workers, and the analysis addresses a situation in which a firm owner F has two incumbent workers, A and B. Replacement workers can be hired at the competitively determined reservation wage $\underline{w}$, which is also the outside payoff to the incumbent workers should they choose to exit. The wages of the incumbents, in contrast, are determined by alternating-offer bargaining with fixed per-period bargaining costs $c_i, i = F, W$, with the incumbent workers making the initial proposals (w_A^0, w_B^0) to the firm in period 0. With respect to each incumbent, the firm can elect to accept (Y) or reject (N) the worker's proposal, or exit (E) and replace that incumbent with a new employee hired at wage $\underline{w}$. If an incumbent's proposal is accepted, the incumbent is paid the proposed wage.

The values of the firm's payoffs are dictated by the given wage proposals and a function $v(n)$ indicating the net value added generated by the employment relationship in terms of the number of incumbent workers $n = 0,1,2$. Assume that the function is strictly increasing in *n*. Thus, the firm's "outside" payoff to replacing a single incumbent while accepting the other incumbent's offer (options YE or EY) is $v(1) - \underline{w} - w_j$, where w_j is the wage paid to the remaining incumbent, and its outside payoff to replacing both incumbents at once is $v(0) - 2\underline{w}$. If the firm accepts both incumbents' wage proposals, its payoff is $v(2) - (w_A + w_B)$.

If the firm rejects just one incumbent's proposal, it makes a counter-offer (w_j^1) to that incumbent in the following period. Let the ensuing subgame beginning with F's initial offer after accepting A's proposal and rejecting B's be denoted $\Gamma^F(YN)$, with corresponding interpretations

for the subgames denoted $\Gamma^F(NY)$, $\Gamma^F(EN)$, and $\Gamma^F(NE)$, noting that all of these subgames are the simpler two-player bargaining games with unique equilibrium solutions discussed earlier. If the firm rejects both initial proposals (option NN), it makes counter-offers (w_A^1, w_B^1) to the two incumbents in the following period. Assuming initially that the incumbents do not collude in determining their responses to these offers, each incumbent simultaneously selects a response Y, N, or E, defined as above.

Parallel to the notation introduced just above, let $\Gamma(YN)$, $\Gamma(NY)$, $\Gamma(NE)$ denote the subgames following the respectively indicated responses from the incumbents, noting again that each of these subgames is a two-person bargaining game with a determinate subgame perfect equilibrium. Finally, let Γ denote the subgame in which A and B make initial proposals, followed by F's responses and the subgames described above, and note that Γ thus represents the entire bargaining game, as well as the subgame encountered in every even-numbered period following the incumbents' simultaneous rejection of the firm's current offers. A detail of the extensive form of the bargaining game just described is shown in Figure 6.

[Figure 6 here]

Consider now the case the case that $c_W < c_F < (v(2) - v(1)) < (v(1) - v(0))$. The first inequality assures that the incumbent workers have some bargaining power; if the inequality were reversed, the firm would get the entire net surplus $v(2) - 2\underline{w}$ regardless of the shape of the value added function. The second inequality simply asserts that the bargaining costs are small relative to the marginal surplus gains from additional incumbents. The final inequality makes it credible for the firm to threaten to fire either incumbent when the incumbents don't collude, and thus makes a "divide and conquer" bargaining strategy viable for the firm in equilibrium. Under these conditions, Skillman and Ryder demonstrate that the equilibrium payoffs are

(25) $2w_I = 2[\underline{w} + (v(2) - (v(1))], \quad \pi_F = 2[v(1) - \underline{w}] - v(2),$

where π_F is the profit received by the firm.

There are two interesting implications of the equilibrium result expressed in (25). First, the firm's equilibrium payoff is strictly decreasing in $v(2)$, the maximum value of the surplus, so that if the firm could choose among alternative production techniques with different value added function, its bargaining payoff would increase by selecting a less efficient technique with a lower value of $v(2)$. Second, this perverse incentive would not exist under a collective bargaining arrangement in which the firm could only choose to replace both incumbents or neither, since in that case (maintaining the other conditions) the firm's payoff is $v(0) - 2\underline{w}$. Generalizing the model to the case in which the firm chooses the size of its labor force, Skillman and Ryder note in addition that the firm's bargaining payoff is not generally maximized at the efficient labor force size n^* that maximizes $v(n) - n\underline{w}$.

There are also, of course, efficiency implications of the bargaining power enjoyed by incumbent workers. As the foregoing analysis illustrates, insider power may go beyond simply counteracting the monopsony power of firms, and drive the wage rates of incumbents above their competitive levels, thus creating unemployment. This aspect of incumbent bargaining power is the focus of the "insider-outsider" theory of Lindbeck and Snower (1989), which provides an alternative to the efficiency-wage account of labor market duality and segmentation.

4.4 Bargaining vs. monopsony and efficiency wage models of labor market outcomes

While the bargaining framework generates some of the same predictions as the efficiency wage model, such as the emergence of supra-competitive wage levels and resulting unemployment or labor market duality, it also offers a basis for unique hypotheses about the economic logic of

wage determination. In particular, bargaining models, unlike the standard efficiency wage or monopsony accounts, suggest that wages will be determined by the total surplus generated by employment relationships, and that the outside payoffs of both firms and employees will influence equilibrium wages. The former result provides a potential explanation for the empirical finding that inter- and intra-industry wage differentials are correlated with industry and firm-level profits (Blanchflower *et al.* 1996, Dickens and Katz 1987, Groshen 1991a).

Machin and Manning (1992) investigate another set of differential implications of the competing theories, concerning the long-run relationship between wages and effort levels. Noting that bargaining, efficiency wage and competitive compensating-differential models all predict a positive long-run or steady-state correlation between wage and effort levels, they demonstrate that the models yield contrasting predictions about the short-run correlation between wages and effort. In particular, the efficiency wage model predicts a positive relationship between current effort and expected future wages (since higher future wages raises the cost of job loss and reduces worker incentives to "shirk"), while a stylized bargaining model predicts an inverse relationship. Testing the alternative models on UK firm-level data, the authors find that the efficiency wage account better explains the data for non-union firms, while the bargaining account provides a better match for unionized firms. Results like theirs argue against taking a "one-theory-fits-all" approach to explaining the operation of labor markets.

5. Imperfect Competition in the Labor Market: The Big Picture

At this point it should be clear what we consider to be the core elements of a suitably comprehensive framework for analyzing imperfect competition in the labor market. While individual researchers will of course want to wield Occam's razor to trim the model to their

particular concerns, the view set out in this chapter suggests that any such model would incorporate the following components in some way:

(i) a model of job and worker flows, including in particular a specification job separations, job matching, and perhaps worker flows into and out of the labor force;

(ii) a wage-setting mechanism that reflects the nature of imperfectly competitive behavior under study, involving unilateral or bilateral wage-setting power; and

(iii) a conception of equilibrium as a (perhaps evolving) steady state involving the balancing of worker and job flows, consisting with optimizing behavior in light of labor market conditions and the wage-setting mechanism.

As illustrated in this chapter's discussion, particular versions of this general framework yield a range of hypotheses about labor market outcomes that are essentially at odds with the predictions of the perfectly competitive model as augmented by human capital theory. These hypotheses offer potential explanations for a range of distinctive labor market phenomena such as unemployment and job queues, discrimination by race and gender, dual labor markets, wage rigidities consistent with the operation of internal labor markets. Implications for appropriate governmental policies vis-à-vis the labor market vary with the set of conditions assumed, however, and are thus ultimately an empirical rather than a theoretical issue.[5]

Endnotes

[1] The discussion in this section is condensed from chapters 5, 7, 18 and 19 in Jacobsen and Skillman (2004).

[2] Though it is worth noting as an empirical matter that it is typical for unemployed workers to receive no offers in a given period; see Clark and Summers (1979).

[3] The discussion in this section is drawn primarily from Jacobsen and Skillman (2004), Chapters 7 and 12, and Osborne and Rubinstein (1990), Chapters 2-4.

[4] A more general analysis of intrafirm bargaining and its implications for organizational design and choice of production technique is provided by Stole and Zwiebel (1996).

References

Akerlof, G. and J. Yellen (1985), 'A near-rational model of the business cycle, with wage and price inertia', *Quarterly Journal of Economics*, **100** (5), 823-38.

Altonji, J.G. and N. Williams (2005), 'Do Wages Rise with Job Seniority? A Reassessment', *Industrial and Labor Relations Review*, **58** (3), 370-97.

Binmore, K., A. Shaked and J. Sutton (1989), 'An outside option experiment', *Quarterly Journal of Economics*, **104** (4), 753-70.

Blanchflower, D.G. and A.J. Oswald (1994), *The Wage Curve*, Cambridge, MA: MIT Press.

Blanchflower, D.G., A.J. Oswald and P. Sanfey (1994), 'Wages, profits, and rent-sharing', *Quarterly Journal of Economics*, **111** (1), 227-51.

Boal, W.M. and M.R. Ransom (1997), 'Monopsony in the labor market', *Journal of Economic Literature*, **35** (1), 86-112.

Boeri, T. and J. van Ours (2008), *The Economics of Imperfect Labor Markets*, Princeton, NJ: Princeton University Press.

Bulow, J. and L. Summers (1986), 'A Theory of Dual Labor Markets with Application to Industrial Policy, Discrimination, and Keynesian Unemployment', *Journal of Labor Economics*, 4 (3), Part 1, 376-414.

Burdett, K. and D.T. Mortensen (1998), 'Wage differentials, employer size, and unemployment', *International Economic Review*, **39** (2), 257-73.

Cain, G.C. (1986), 'The economics of labor market discrimination', in *Handbook of Labor Economics,* **2**, Ashenfelter, O. and R. Layard (eds), Amsterdam and New York: North-Holland, pp. 693-785.

Clark, K.B. and L.H. Summers (1979), 'Labor market dynamics and unemployment: a reconsideration', *Brookings Papers on Economic Activity*, **1979** (1), 13-60.

Davis, S.J., J.C. Haltiwanger and S. Schuh (1996), *Job Creation and Destruction*, Cambridge, MA: MIT Press.

Dickens, W. and K. Lang (1993), 'Labor market segmentation theory: reconsidering the evidence', in *Labor Economics*: *Problems in Analyzing Labor Markets*., Darity, W. Jr. (ed.), Norwell, MA: Kluwer Academic Publishers, pp. 141-80.

Dickens, W. and L. Katz (1987), 'Interindustry wage differences and industry characteristics', in *Unemployment and the Structure of Labor Markets*, Lang, K. and J. Leonard (eds), Oxford, UK: Basil Blackwell, pp. 48-89.

Doeringer, P.B. and M.J. Piore (1971), *Internal Labor Markets and Manpower Analysis*, Armonk, NY: M.E. Sharpe.

Dorman, P. (1996), *Economics, Dangerous Work, and the Value of Human Life*, Cambridge, UK: Cambridge University Press.

Groshen, E. (1991a), 'Five reasons why wages vary among employers', *Industrial Relations*, **30** (3), 350-81.

Groshen, E. (1991b), 'Sources of Intra-Industry Wage Dispersion: How Much Do Employers Matter?', *Quarterly Journal of Economics*, **106** (3), 869-84.

Jacobsen, J.P. (2007), *The Economics of Gender, Third Edition*, Malden, MA and Oxford, UK: Blackwell Publishing.

Jacobsen, J.P. and G.L. Skillman (2004), *Labor Markets and Employment Relationships*: *A Comprehensive Approach*, Malden, MA and Oxford, UK: Blackwell Publishing.

Kerr, C. (1988), 'The neoclassical revisionists in labor economics (1940-1960)—R.I.P.', in *How Labor Markets Work*, B.E. Kaufman (ed.), Lexington, MA: D.C. Heath, pp. 1-46.

Krueger, A. and L. Summers (1988), 'Efficiency wages and the inter-industry wage structure,' *Econometrica*, **56** (2), 259-93.
Laffont, J. and D. Martimort (2001), *The Theory of Incentives*: *The Principal-Agent Model*, Princeton, NJ: Princeton University Press.
Lindbeck, A. and D.J. Snower (1989), *The Insider-Outsider Theory of Employment and Unemployment*, Cambridge, MA: MIT Press.
Machin, S. and A. Manning (1992), 'Testing dynamic models of worker effort', *Journal of Labor Economics*, **10** (3), 288-305.
Manning, A. (2003), *Monopsony in Motion*: *Imperfect Competition in Labor Markets*, Princeton, NJ: Princeton University Press.
McCall, B. and J. McCall (2007), *The Economics of Search*, New York: Routledge.
Osborne, M.J. and A. Rubinstein (1990), *Bargaining and Markets*, San Diego, CA: Academic Press.
Osterman, P. (ed.) (1984), *Internal Labor Markets*, Cambridge, MA: MIT Press.
Rubinstein, A. (1982), 'Perfect equilibrium in a bargaining model', *Econometrica*, **50** (1), 97-109.
Saint-Paul, G. (1996), *Dual Labor Markets*: *A Macroeconomic Perspective*, Cambridge, MA: MIT Press.
Scaramozzino, P. (1991) 'Bargaining with outside options: wages and employment in US manufacturing', *Economic Journal*, **101** (405), 331-42.
Skillman, G.L. and H.E. Ryder (1993), 'Wage bargaining and the choice of production technique in capitalist firms', in *Markets and Democracy*: *Participation, Accountability*, *and Efficiency*, Bowles, S., H. Gintis and B. Gustafsson (eds), Cambridge, UK: Cambridge University Press, pp. 217-27.
Stole, L.A. and J. Zwiebel (1996), 'Organizational design and technological change under intrafirm bargaining', *American Economic Review*, **86** (1), 195-222.
Troske, K. (1999), 'Evidence on the employer-size wage premium from worker-establishment matched data', *Review of Economics and Statistics*, **81** (1), 15-26.
Weiss, A. (1990), *Efficiency Wages*: *Models of Unemployment*, *Layoffs, and Wage Dispersion*, Princeton, NJ: Princeton University Press.

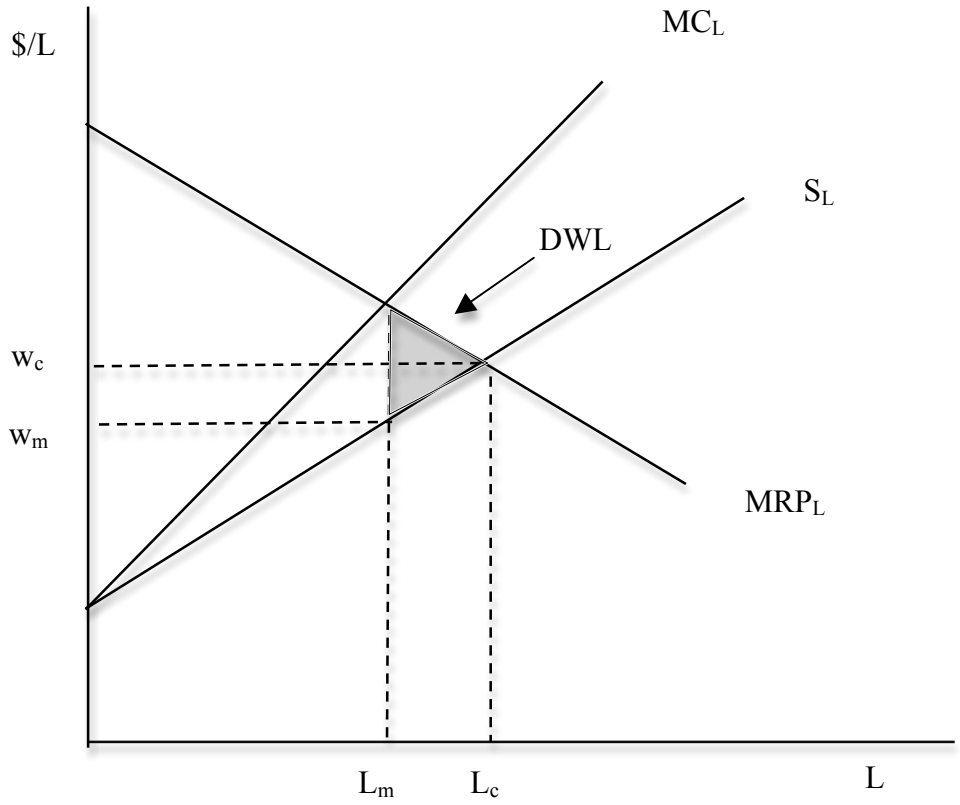

Figure 1
Pure monopsony equilibrium

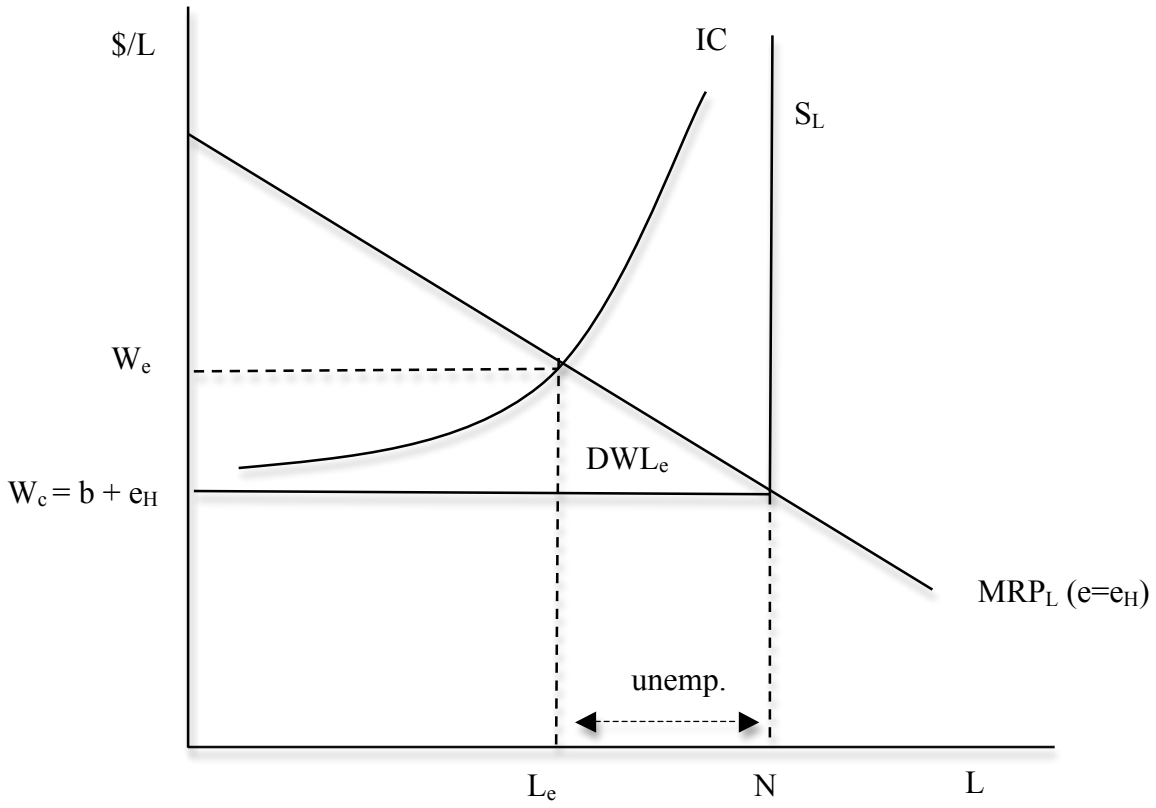

Figure 2
Efficiency wage equilibrium

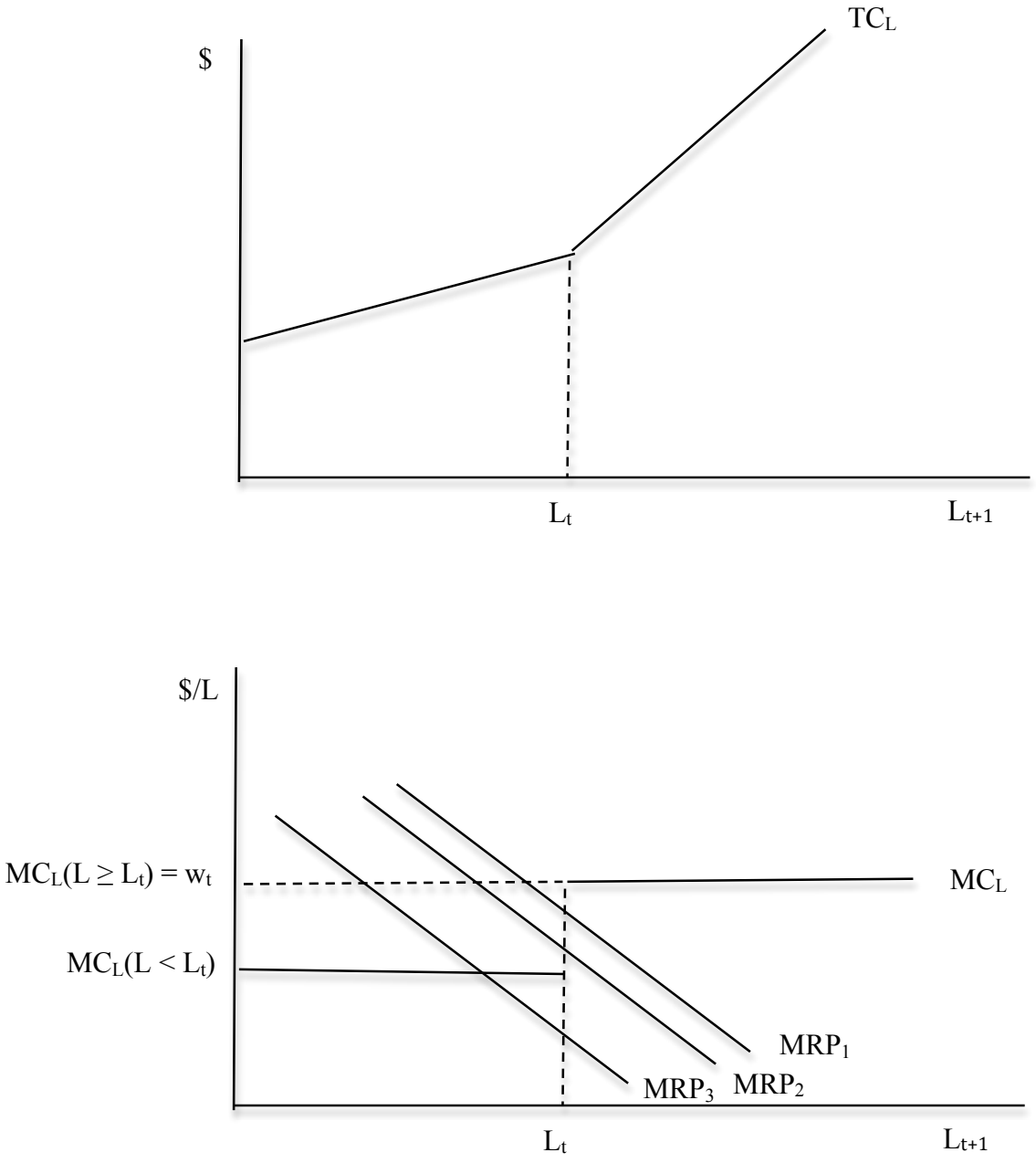

Figure 3
Wage rigidity and hysteresis

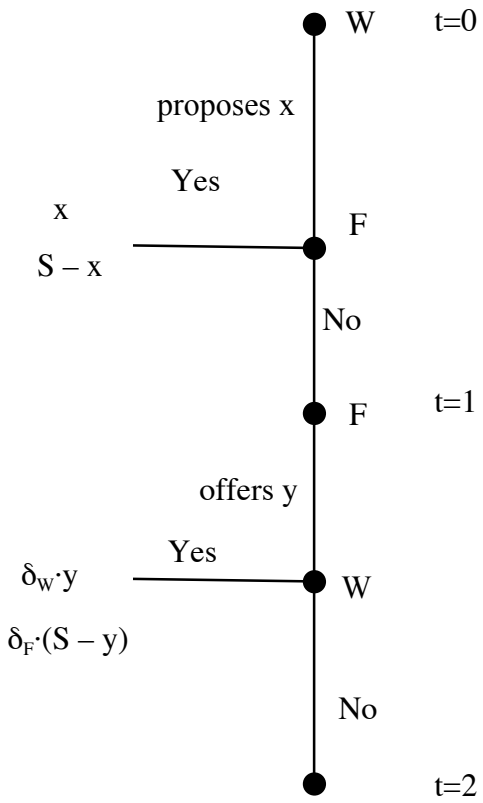

Figure 4
A detail of the alternating-offers bargaining game

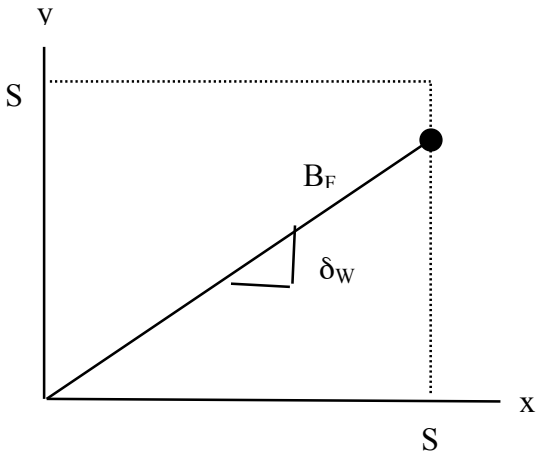

Figure 5A
The firm's bargaining reaction function B_F

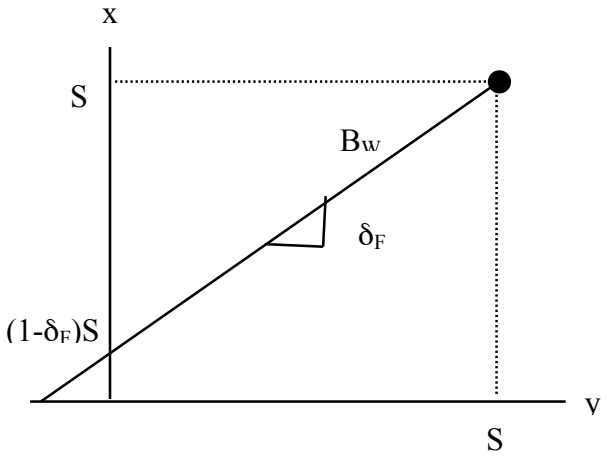

Figure 5B
The worker's bargaining reaction function B_W

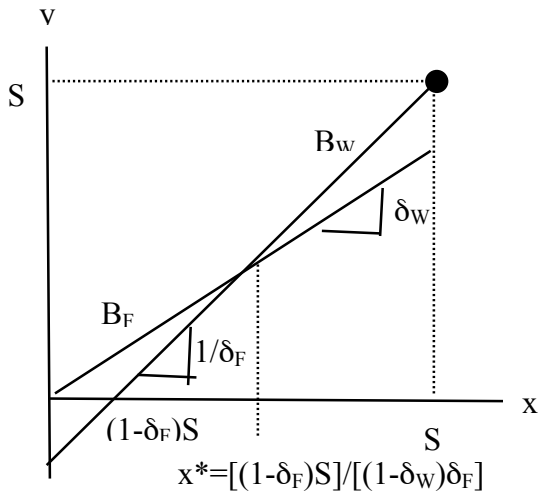

Figure 5C
Rubinstein bargaining equilibrium with discounting and no exit options

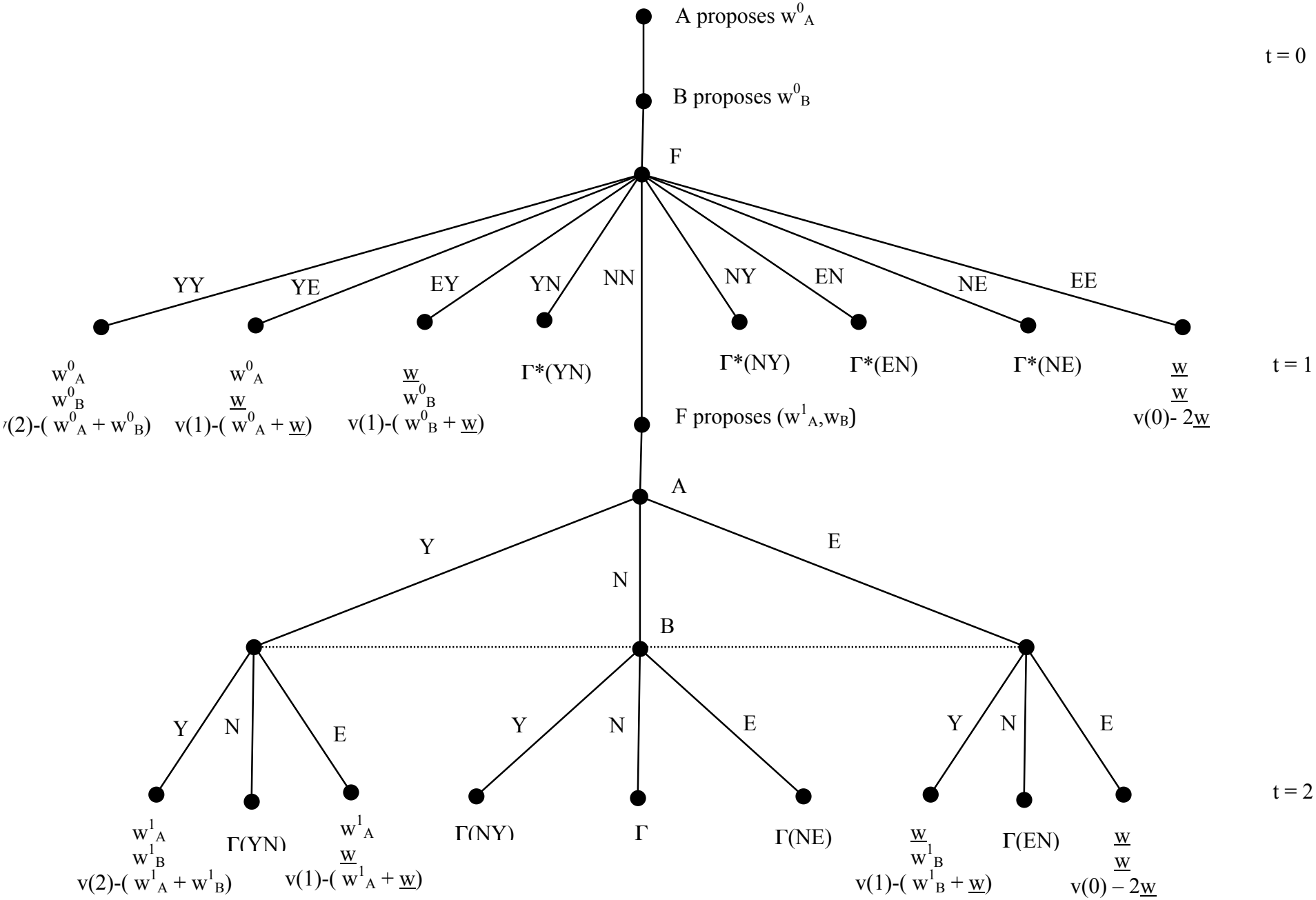

Figure 6
A detail of the multilateral bargaining game