European Economic Review 46 (2002) 1733 – 1756
www.elsevier.com/locate/econbase

The labor market and corporate structure

Daron Acemoglu[a,*], Andrew F. Newman[b]

[a]Department of Economics, Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, MA 02142-1347, USA
[b]Department of Economics, University College London, London, UK

Received 1 May 2000; accepted 2 October 2001

Abstract

We analyze the impact of labor demand and labor market regulations on the corporate structure of firms. Higher wages are associated with lower monitoring, irrespective of whether these high wages are caused by labor market regulations, unions or higher labor demand. We also find that the organization of firms has important macroeconomic implications. In particular, monitoring is a type of "rent-seeking" activity and the decentralized equilibrium spends excessive resources on monitoring. Labor market regulations that reduce monitoring by pushing wages up may increase net output or reduce it only by a small amount even though they reduce employment. © 2002 Elsevier Science B.V. All rights reserved.

JEL classification: J41; L23.

Keywords: Corporate structure; Efficiency wages; Labor market regulations; Monitoring; Moral hazard

1. Introduction

The way firms are organized to provide incentives to their employees varies across countries and changes over time. Fig. 1 plots one measure of organizational form related to the extent of monitoring; the ratio of managerial to production workers in six countries.[1] The U.S. and Canada have more managers per worker than the other

* Corresponding author. Tel.: +1-617-253-1927; fax: +1-617-253-1330.
E-mail address: daron@mit.edu (D. Acemoglu).

[1] The data plotted are ratios of managers to non-agricultural, non-managerial workers, calculated from the Labor Statistics of the International Labor Organization. The sources of the data are the labor force surveys of the respective countries (Current Population Survey for the U.S.). These numbers have to be interpreted carefully because the definition of a manager may vary across countries, and may even change over time. Nevertheless, this is the only data set that we are aware of with cross-country variations in the number of supervisory employees. Other countries do not have enough years to construct a time-series in the ILO data

0014-2921/02/$ - see front matter
PII: S0014-2921(01)00199-4

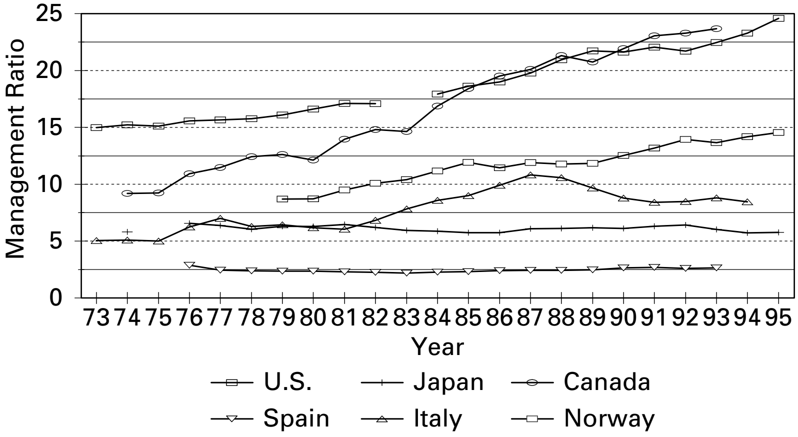

Fig. 1. Trends in the ratio of managerial employees to non-managerial, non-agricultural workers in six countries. *Source*: ILO Labor Statistics.

countries. Moreover, while the ratio of managers to workers is constant or increasing only slightly in Italy, Spain, Japan and Norway, it appears to increase rapidly in the U.S. and Canada. A large industrial relations and business history literature also emphasizes international and temporal variations in business practices and organizations. [2]

What could account for these differences in organizational forms? Although differences in technology could account for these patterns, it is plausible that all OECD countries have access to the same technological possibilities frontier. [3] In this paper we take an alternative approach and argue that labor demand and labor market regulations lead to endogenous differences in organizational forms. Our main focus is on "the corporate structure", which we take to be related to the extent and form of monitoring.

Corporate structure itself is a choice variable for firms, and like many of their decisions, will be determined partly by market conditions. We construct a simple general equilibrium model in which conditions in the labor market—both supply-and-demand and regulatory—will lead firms to make different organizational choices. Moreover, we will find that firms' responses to changes in labor market fundamentals are in line with the time trends indicated in Fig. 1. We will also show that the organizational choices of firms not only respond to the state of the macroeconomy, but can also have a substantial influence on its performance. This is because firms spend considerable resources on monitoring, which could otherwise be used for directly productive

set, but cross-sectional comparisons are in line with the data reported here: with the exception of the U.K. all other countries appear to have substantially lower ratios of managerial workers in their workforces than the U.S. and Canada.

[2] See, for example, Appelbaum and Batt (1994) for an overview, Chandler (1977) for a history of U.S. frms, and Freeman and Lazear (1994) on the contrast of some aspects of U.S. and German labor relations.

[3] Another possibility is that which workers count as managers differs across countries. While this is undoubtedly true, it does not explain the differential trends across countries.

activities. Therefore, it is important to take the organizational implications of labor market policies into account in calculating their welfare consequences.

The logic of our approach is best understood by considering the incentives of a single worker. After signing on with the firm, the worker takes an unobserved effort decision: he may work or shirk. Although this effort choice is not directly observed, the firm can detect it with a certain probability that depends on the amount of resources devoted to monitoring.[4] The contract between worker and firm will specify a compensation level which depends on whether he has been detected shirking. Crucially, we assume that there are liability limitations on workers: no contract can punish a worker arbitrarily severely. As is well-known, this will lead to equilibrium rents (efficiency-wages) for workers in order to induce them to exert effort.[5]

In making his effort decision, the worker takes account of three factors: (i) the wage (rent) he is risking to lose by shirking; (ii) his payoff if fired for having shirked; and (iii) the probability of being detected when shirking (see for example Shapiro and Stiglitz, 1984; Calvo and Wellisz, 1979). The key point is that all three will be affected not merely by the technology of the firm or legal contractual restrictions, *but also by market conditions*. The main market condition that we will focus on in this paper is the alternative employment opportunities facing the worker, which are themselves determined by the state of labor demand and labor market regulations.

Consider the impact of labor demand on incentives and monitoring. An increase in labor demand creates three effects on the worker's incentives corresponding to three factors that the worker takes into account in making his effort decision. The first is the *ex-ante utility effect*: in a tighter labor market, the ex ante utility and the equilibrium wages of workers are higher because firms are competing in order to attract workers.[6] When limited liability constraints prevent negative wages, a high level of compensation naturally translates into high powered incentives. In other words, when firms are forced to pay high wages to workers because of market conditions, they can also use these attractive wages to provide them with the right incentives and do not need a high level of monitoring.

The second force is the *ex-post reservation utility effect*. When labor demand is high workers know that being fired is not a harsh punishment because they can get a new job relatively easily (Shapiro and Stiglitz, 1984). This implies that firms will need to monitor their employees closely when labor demand is high.

Finally, in a tight labor market, the demand for the resources used for monitoring will also increase. For example, when workers are used to monitor other workers, the cost of monitoring will increase with the level wages. This *cost-of-monitoring effect* also works in the direction of reducing monitoring in tight labor markets: when the

[4] In this context, monitoring should be interpreted broadly: anything which provides some information about worker effort is valuable to the firm (Holmstrom, 1979).

[5] In practice, firms have access to a variety of other tools to provide the correct incentives to workers, including tournaments and bonus pay (e.g., Gibbons, 1996; Prendergast, 1999). However, when workers cannot be paid negative wages, the employment relation has to offer them some amount of rent.

[6] The firm is offering a high ex ante utility, rather than an ex-post utility, because even if the worker is hired with an attractive contract, he will receive a low wage and will be fired if he is caught shirking. This distinction between ex-ante and ex-post values will be important in our analysis.

cost of monitoring is high as in times of buoyant labor demand, firms will want to use less of it. We will show that the first and the third effects always dominate the second: when higher labor demand increases wages, the amount monitoring is reduced.

Another set of variables that vary across countries and time periods is labor market regulations and institutions. In particular, unions and minimum wage type regulations increase wages relative to labor productivity. Again the ex-ante utility effect comes into action; since firms are paying high wages, they can also use these to provide them with the right incentives and so reduce monitoring. [7]

One reason to be interested in the choice of corporate structure is that it has important macroeconomic implications. The basic result we obtain here is that the decentralized equilibrium spends excessive resources on monitoring, and since these resources could have been used more productively, it fails to maximize net output. The intuitive reason is that monitoring is at some level a type of "rent-seeking" activity: it enables the firm to reduce wages, transferring resources from workers to firms. A social planner who cares only about aggregate output would want to raise payments to workers in order to save on monitoring costs.

This simple model also offers a possible interpretation for the cross-country trends shown in Fig. 1. If, as many economists believe, wages are high in Europe because of labor market regulations and more powerful unions, then our model implies that European firms should spend less on monitoring. Moreover, since during the past three decades wages for unskilled and production workers in the U.S. have fallen by as much as 30% (e.g. Freeman, 1995), our model suggests that this period should also experience an increase in the extent of monitoring to restore worker incentives eroded by falling wages.

Our work is clearly related to the "efficiency wage" literature of a decade ago (Foster and Wan, 1984; Bulow and Summers, 1986; Rebitzer and Taylor, 1991; see Katz, 1987, and Weiss, 1990, for surveys), especially to the work of Shapiro and Stiglitz (1984). The main difference of our analysis is that we endogenize the monitoring technology and try to understand the cross-country patterns and temporal developments in the organization of firms through this lens. Calvo and Wellisz (1979) also endogenize monitoring in an efficiency wage model but without our focus on the general equilibrium determinants of the extent of monitoring. To our knowledge, ours is the only model that analyzes the impact of labor demand and regulations on corporate structure and economic performance. More generally, although there are other frameworks for analyzing the internal organization of the firm, we find a model with efficiency wages and endogenous monitoring the natural place to start for an analysis of external factors on the internal organization of the firm. Gordon (1996) has also pointed out some of the same differences between the U.S. and other economies' corporate organizations, but sought to explain these differences by arguing that corporate bureaucracies have a tendency to expand, and they have been allowed to do so in the U.S. Finally, our paper is also related to studies that analyze the effect of the technical progress and/or

[7] The exception of course is regulations that directly or indirectly prevent firms from firing workers that shirk.

demographic changes on organizational changes and wage inequality, for example, Acemoglu (1999), Caroli and van Reenen (1999), and Kremer and Maskin (1998).

Our model in Section 2 is a static one which only focuses on the ex-ante utility effect we mentioned above. We analyze this model in detail because it contains most of the economics in this paper. In Section 3, we generalize our model to a dynamic setting which also incorporates the ex-post utility and cost-of-monitoring effects discussed above. The set-up is based on Shapiro and Stiglitz's model, but nests their model as well as our model of Section 2 as special cases. We show that in the original Shapiro–Stiglitz model, labor market regulations that increase wages will have the same effect as in our static model, but a change in labor demand conditions would leave the degree of monitoring unchanged. This is because the trade-off between wages and monitoring is such that firms prefer to increase wages but leave monitoring unaffected in response to a tightening labor market. We demonstrate, however, that this result is not robust. For example, if firms can have contractual arrangements with their workers (as in our static model), or if the cost of monitoring is endogenized so that it changes with the state of labor demand, then a tighter labor market will lead to less monitoring. The paper concludes with a brief extension of our basic framework to discuss income distribution, and a short appendix contains some technical proofs.

2. A static model

We start with a one-period model which illustrates the basic ideas in the simplest environment. In particular, it focuses on the ex-ante utility effect, and abstracts from the other two effects discussed in the introduction. Those will be incorporated in the dynamic model of the next section and shown not to affect the basic qualitative conclusions reached with the static model.

2.1. Basics

Consider a one-period economy consisting of a continuum of measure N of workers and a continuum of measure 1 of firm owners who are different from the workers. [8] Each firm i has the production function $AF(L_i)$ where L_i is the number of workers it hires who choose to exert effort. Workers who shirk are not productive. We also assume that firms are large, so that the output of an individual worker is not observable and therefore not contractible.

A firm can use other information to give the correct incentives to its workers. A worker's actions affect the probability distribution of some observable signal on the

[8] We are implicitly assuming that firms own some capital (human or physical) which the workers do not. This prevents free entry which would compete away all firm profits (which are really just returns to this capital). The "limited liability" assumption we make below simultaneously helps to explain why workers cannot post bonds (which would obviate the need for monitoring), why they have imperfect access to capital, and why the power of incentives increases in the level of compensation. For a model which explicitly treats the role of capital market imperfections in determining the type and efficiency of organizational form, see Legros and Newman (1996).

basis of which the firm compensates him (e.g. Holmstrom, 1979). Specifically, when the worker exerts effort, this signal takes the value 1. When he shirks, this signal is equal to 1 with probability $1-q_i$ and 0 with probability q_i. The worker, like the firm, is risk-neutral and maximizes income minus effort cost which is denoted by e.

The probability of detecting low effort by the worker, q_i, is determined by a host of factors including the production technology used by the firm, the numbers of supervisors, managers and accountants, and more generally the information technology of the firm (e.g. computers and cameras). Our analysis turns on the fact that firms are able to choose many of these attributes of organization. We model this by allowing q_i be a decision variable for firms. We assume that $q_i = q(m_i)$ where m_i is the degree of monitoring per worker by firm i; the cost of monitoring for firm i which hires L_i workers is sm_iL_i. For example, we can think of m_i as the number of managers per production worker and s as the salary of managers. For now, s is fixed and exogenous. In the next section, we will endogenize s as an equilibrium outcome. We assume that q is increasing, concave and differentiable with $q(0)=0$ and $q(m)<1$ for all m. The choice of q_i in our model will be the crucial aspect of organizational form.[9]

Since there is a limited liability constraint, workers cannot be paid a negative wage, and the worst thing that can happen to a worker is to receive zero income. Since all agents are risk-neutral, without loss of generality we can restrict attention to the case where workers are paid zero when caught shirking. Therefore, the incentive compatibility constraint of a worker employed in firm i can be written as: $w_i - e \geqslant (1-q_i)w_i$. If the worker exerts effort, he gets utility $w_i - e$, which gives the left-hand side of the expression. If he chooses to shirk, he gets caught with probability q_i and receives zero. If he is not caught, he gets w_i without suffering the cost of effort. This gives the right-hand side of the expression. In writing this expression, we have assumed that a worker who gets fired from his job does not receive z, but this is of no major consequence for our results.

Firm i's maximization problem can be written as

$$\max_{w_i,L_i,q_i} \Pi = AF(L_i) - w_iL_i - sm_iL_i \tag{1}$$

subject to

$$w_i \geqslant \frac{e}{q(m_i)}, \tag{2}$$

$$w_i - e \geqslant \underline{u}. \tag{3}$$

The first constraint is the incentive compatibility condition rearranged. The second is the participation constraint where $\underline{u}$ is the ex ante reservation utility (outside option) of the worker; in other words, what he could receive from another firm in this market.

It is important to bear in mind the difference between the ex ante and ex post outside options. These play distinct roles in the worker's incentive problem, and are affected differently by market conditions. Specifically, if the worker gets fired for shirking, he does *not* receive $\underline{u}$ but instead gets 0, his ex post outside option (recall there is no

[9] In fact, throughout most of the paper q will be the only endogenous aspect of organizational form; in Section 5 we will discuss some other dimensions of corporate structure.

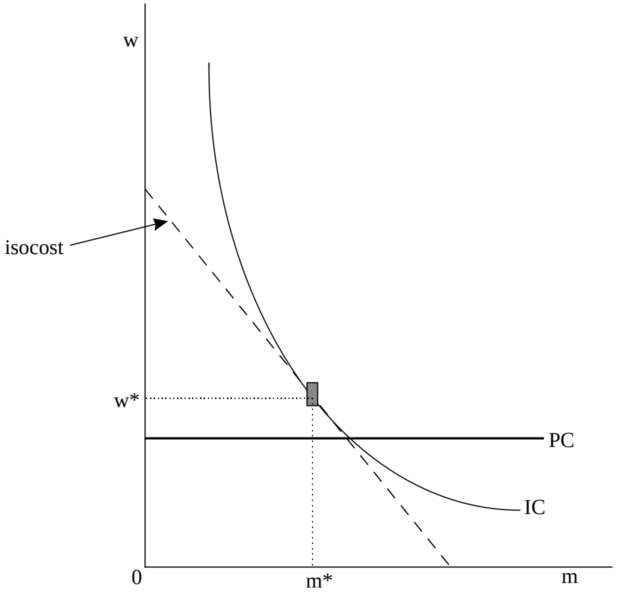

Fig. 2. Participation constraint is slack.

more production after the first period). On the other hand, the firm takes the ex ante reservation utility $\underline{u}$ as given: Constraint (3) reflects the fact that it is not enough for a firm to convince the worker to exert effort once he has joined the firm; it also has to convince them to join the firm in the first place.

The maximization problem (1) has a recursive structure: m and w can be determined first without reference to L by minimizing the cost of a worker $w + sm$ subject to (2) and (3); then, once this cost is determined, the profit maximizing level of employment can be found (the recursiveness of this problem is similar to that in Calvo and Wellisz (1979)). Each subproblem is strictly convex, so the solution is uniquely determined, and all firms will make the same choices: $m_i = m$, $w_i = w$ and $L_i = L$. In other words, the equilibrium will be symmetric.

Another useful observation is:

Lemma 1. In equilibrium, *the incentive compatibility constraint*, (2), *always binds.*

If the incentive compatibility constraint, (2), did not bind, the firm could lower q, and increase profits without affecting anything else. This differs from the simplest moral hazard problem with fixed q in which the incentive compatibility constraint (2) could be slack.

By contrast, the participation constraint (3) may or may not bind. The comparative statics of the solution have a very different character depending on whether it does. The two situations are sketched in Figs. 2 and 3. When (3) does not bind, the solution is characterized by the tangency of the (2) with the per-worker cost $w + sm$ (Fig. 2).

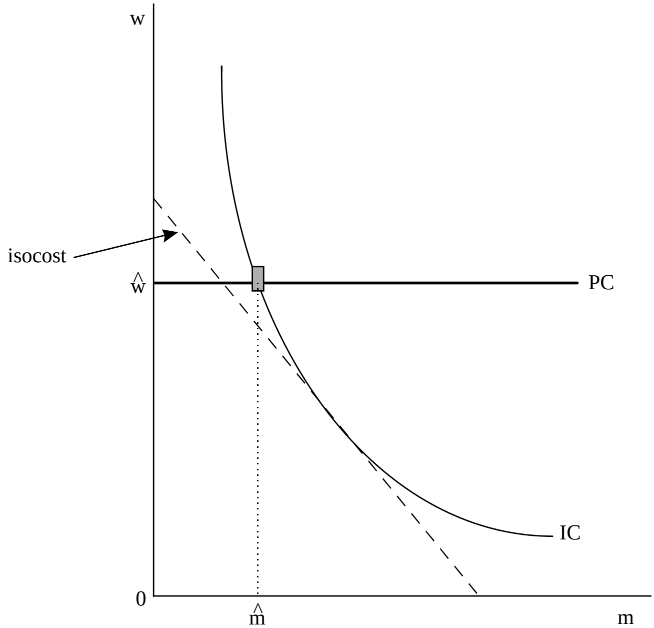

Fig. 3. Participation constraint is binding.

Call this solution (w^*, m^*), where

$$\frac{eq'(m^*)}{(q(m^*))^2} = s \quad \text{and} \quad w^* = \frac{e}{q(m^*)}. \tag{4}$$

In this case, because the participation constraint (3) does not bind, w and m are given by (4) and small changes in $\underline{u}$ leave these variables unchanged. In contrast, if (3) binds, w is determined directly from this constraint as equal to $\underline{u} + e$, and an increase in $\underline{u}$ causes the firm to raise this wage. Since (2) holds in this case, the firm will also reduce the amount of information gathering, m.

What determines whether (3) binds? Let $\hat{w}$ and $\hat{m}$ be the per-worker cost minimizing wage and monitoring levels (which would not be equal to w^* and m^* when (3) binds). Then, labor demand of a representative firm solves

$$AF'(\hat{L}) = \hat{w} + s\hat{m}. \tag{5}$$

Using labor demand, we can determine $\underline{u}$, workers' ex ante reservation utility from market equilibrium. It depends on how many jobs there are. If aggregate demand $\hat{L}$ is greater than or equal to N, then a worker who turns down a job is sure to get another. In contrast, if aggregate demand $\hat{L}$ is less than N, then a worker who turns down a job may end up without another. In particular, in this case, $\underline{u} = (\hat{L}/N)(\hat{w} - e) + (1 - (\hat{L}/N))z$,

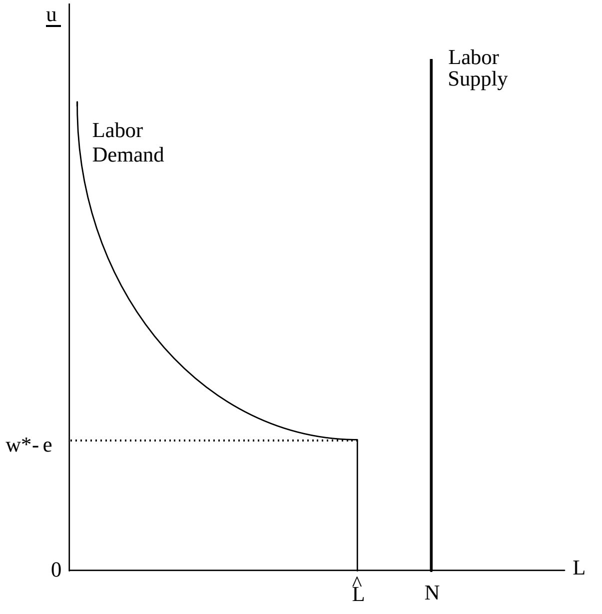

Fig. 4. Participation constraint is slack.

where z is an unemployment benefit that a worker who cannot find a job receives.[10] The unemployment benefit z will be useful for some of our comparative statics below, and we always assume that z is not large enough to shut down the economy.

When $\hat{L} = N$, there are always firms who want to hire an unemployed worker at the beginning of the period, and thus $\underline{u} = \hat{w} - e$. If there is excess supply of workers, i.e. $\hat{L} < N$, then firms can set the wage as low as they want, and so they will choose the profit maximizing wage level w^* as given by (4). In contrast, with full employment, firms have to pay a wage equal to $\underline{u} + e$ which will generically exceed the (unconstrained) profit maximizing wage rate w^*. Therefore, we can think of labor demand as a function of $\underline{u}$, the reservation utility of workers: Firms are "utility-takers" rather than price-takers. Figs. 4 and 5 show the two cases; the outcome depends on the state of labor demand. More important, the comparative statics are very different in the two cases.

More formally, an equilibrium in this economy is a vector $(\underline{u}, \hat{w}, \hat{m}, \hat{L})$ such that (i) given $\underline{u}$, $(\hat{w}, \hat{m}, \hat{L})$ are chosen to maximize (1) subject to (2) and (3); (ii) $\hat{L} \leqslant N$;

[10] Recall that $\underline{u}$ is the ex ante reservation utility, and loosely speaking, it refers to what the worker could get before the market clears. In particular, if the worker does not agree with the firm that he is in contact with, he is sure to find a new firm when there is excess demand. In contrast, when $\hat{L} < N$, we think of disagreement as the worker taking another draw from the distribution, and he will have a probability $\hat{L}/N$ of ending up with a firm. It might be thought that this probability should be less than $\hat{L}/N$, since some of the workers and firms may have agreed already. This would not, however, change any of the results, because in the case where $\hat{L} < N$, the participation constraint is slack anyway.

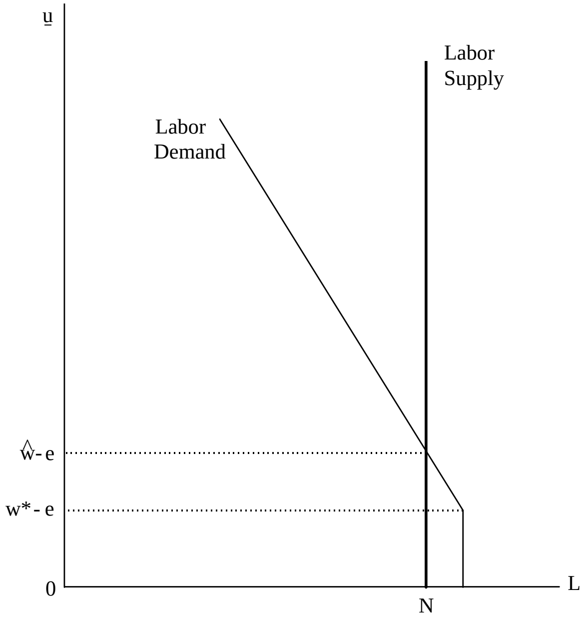

Fig. 5. Participation constraint is binding.

and (iii) $\underline{u} = z + \min\{1, \hat{L}/N\}(\hat{w} - e - z)$. Note that workers' reservation utility level, $\underline{u}$, plays the role of a price in equilibrating the market.

Proposition 2. An equilibrium exists, is unique and takes one of two forms.

1. *Full employment equilibrium* (*FEE*) *in which* (3) *holds as an equality, thus* $\underline{u} = \hat{w} - e$.
2. *Unemployment equilibrium* (*UE*) *in which* (3) *is slack and thus* $\underline{u} < \hat{w} - e$ *and* $\hat{w} = w^*$ *and* $\hat{m} = m^*$ *as given by* (4).

The proof of this result is straightforward and is omitted; inspection of Figs. 4 and 5 should suffice to make it plausible. In FEE, the participation constraint (3) binds, $\hat{w} > w^*$ and $\hat{m} < m^*$. In this case, the market forces firms to pay wages higher than their unconstrained optimum w^*; as a consequence, they engage in less than their privately optimal level of monitoring m^*. By contrast, in UE when (3) is slack, there is an "excess supply" of workers, and firms choose (w^*, m^*). Cutting wages below w^* would still attract workers, but would be unprofitable for firms because in order to prevent shirking, they would have to increase m above their optimum m^*.

2.2. Comparative statics

First, consider a small increase A and suppose that (3) is slack. The tangency between (2) and the per worker cost, shown in Fig. 2, is unaffected. Therefore, neither w nor m change. Instead, the demand for labor in Fig. 4 shifts to the right and firms hire more workers. As long as (3) is slack (that is as long as the vertical portion of labor demand remains to the left of N in Fig. 4), firms will continue to choose their (market) unconstrained optimum, (w^*, m^*), which is independent of the marginal product of labor. As a result, changes in labor demand do not affect the organizational form of the firm.

If instead (3) holds as an equality, comparative static results will be different. In this case, (2), (5), and $L = N$ jointly determine $\hat{q}$ and $\hat{w}$. An increase in A induces firms to demand more labor, increasing $\hat{w}$. Since (2) holds, this reduces $\hat{q}$ as can be seen by shifting the PC curve up in Fig. 5. Therefore, when (3) holds, an improvement in the state of labor demand reduces monitoring. The intuition is closely related to the fact that workers are subject to limited liability. When workers cannot be paid negative amounts, the level of their wages is directly related to the *power* of the incentives. The higher are their wages, the more they have to lose by being fired and thus the less willing they are to shirk.

Next, suppose that the government introduces a wage floor $\underline{w}$ above the equilibrium wage (or alternatively, unions demand a higher wage than would have prevailed in the non-unionized economy). It is straightforward to see that Lemma 1 still holds so that the incentive compatibility constraint (2) will never be slack. Therefore, a higher wage will simply move firms along the IC curve in Fig. 3a and reduce m. However, this will also increase total cost of hiring a worker, reducing employment.

Finally, suppose that z changes and that $\underline{u} > z$. Then if (3) is slack, a small change in z affects neither m nor w. On the other hand, when $\underline{u} = z$, a rise in z increases w and reduces m. Also it is important to note that both high z and high $\underline{w}$ make an unemployment equilibrium more likely than a full employment equilibrium, whereas a high level of labor productivity A makes a full employment equilibrium more likely.

These comparative static results suggest a possible interpretation for the differences in the corporate structure in the U.S. and Europe. European economies, characterized by high minimum wages and unemployment benefits, [11] are more likely to be in the unemployment equilibrium, and thus our model suggests that they should have less monitoring and a lower ratio of managerial to production workers. The comparative statics with respect to A, in turn, suggest that a change in the demand for production workers (say due to technical change or international trade) should have a very different impact in an economy in the full employment equilibrium as compared to an economy in the unemployment equilibrium. If we think of the U.S. as in the full employment regime and the more regulated European economies as in the unemployment regime, our simple model predicts that in response to a falling demand for unskilled and semi-skilled workers, wages should fall in the U.S. and the degree of monitoring should increase. In contrast, in Europe, only unemployment rates should increase.

[11] Another labor market regulation that is common in Europe is severance pay (firing costs). These are not as straightforward to incorporate into our model. At one level, they would act similar to an increase to z, but they would also make firing less desirable for firms and perhaps make the threat of firing less credible. This will tend to weaken worker incentives, which will tend to raise monitoring levels and be unambiguously detrimental to aggregate performance.

2.3. Welfare

Consider the aggregate surplus Y generated by the economy:

$$Y = AF(L) - smL - eL, \tag{6}$$

where $AF(L)$ is total output, and eL and smL are the (social) input costs.

In this economy, the equilibrium is constrained Pareto efficient: subject to the informational constraints, a social planner could not increase the utility of workers without hurting the owners. But total surplus Y is never maximized in laissez-faire equilibrium:

Proposition 3. The decentralized equilibrium never maximizes Y. Subsidizing w and taxing profits would increase Y.

This result follows from noting that if we can reduce q without changing L, then Y increases. A tax on profits used to subsidize w relaxes the incentive constraint (2) and allows a reduction in monitoring.[12] Indeed, the second-best allocation which maximizes Y subject to (2) would set wages as high as possible subject to zero profits for firms. Suppose that the second-best optimal level of employment is $\tilde{L}$, then we have[13]

$$\tilde{w} + sq^{-1}\left(\frac{e}{\tilde{w}}\right) = \frac{AF(\tilde{L})}{\tilde{L}}. \tag{7}$$

In this allocation, all firms would be making zero-profits; since in the decentralized allocation, due to decreasing returns, they are always making positive profits, the two will never coincide.

A different intuition for why the decentralized equilibrium fails to maximize net output is as follows: part of the expenditure on monitoring, smL, can be interpreted as "rent-seeking" by firms. Firms are expending resources to reduce wages—they are trying to minimize the private cost of a worker $w + sm$—which is to a first-order approximation, a pure transfer from workers to firms. A social planner who cares only about the size of the national product wants to minimize $e + sm$, and therefore would spend less on monitoring. Reducing monitoring starting from the decentralized equilibrium would therefore increase net output.

Fig. 6 draws the equilibrium, first-best and second-best surpluses as a function of the supply of labor N for a parametric case. Over the range of N depicted, the first-best, which prevails when moral hazard is absent, is simply given by $AF(N) - Ne$. The second-best adopts the wage rule (7) and also chooses full employment in this range. Finally, the equilibrium is the outcome characterized in Proposition 2. Observe that, over a certain range, the equilibrium surplus is decreasing in N. This is because high levels of N reduce wages through the usual supply effect and thus induce firms to

[12] However, recall that w is the wage that workers receive only when they are not caught shirking. Subsidizing wages irrespective of whether workers are caught shirking or not would not affect the incentive compatibility constraint and would not have this beneficial effect.

[13] Since the social cost of a worker is $e + sm$, as long as m and s cannot be made equal to zero, the optimal level of employment will be lower than it would be under the full-information first-best. However, since the planner is minimizing this social cost, he will always want to employ at least as many workers as the decentralized equilibrium would.

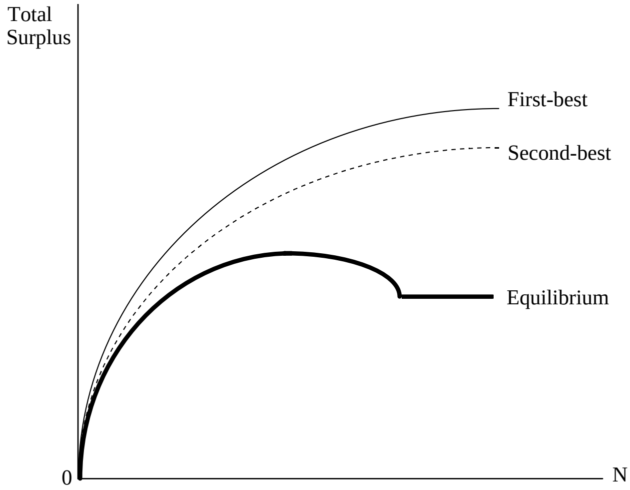

Fig. 6. Plot of the total surplus in first-best, second-best and equilibrium.

increase m in order to ensure incentive compatibility. This suggests that over a certain range labor market policies which increase wages and reduce employment will actually increase surplus, or at least have only a small effect on aggregate welfare. This raises the possibility that economies with very different allocations in terms of employment, organizational form and (functional) distribution of income (such as the U.S. vs. continental European economies) may achieve similar levels of total output or efficiency.

3. A dynamic model

We now analyze a dynamic model that generalizes the main results of the previous section by incorporating some of the effects discussed in the Introduction, which were missing from the static model. Most of the results of the static case generalize to this dynamic environment, but additionally, even in the equivalent of "unemployment regime", an increase in the productivity of production workers will lead to higher wages and less monitoring, thus to a different corporate structure.

3.1. The environment

As before, there is a measure N of workers and a unit measure of firms. Time is continuous. All agents are risk-neutral, infinitely-lived, and discount the future at the rate r. Workers can be employed either to produce output or to supervise other workers. If a supervisor (manager) is monitoring $1/m$ workers, then a shirking worker is detected with probability $q(m)$. As before, workers are never mistakenly detected shirking. The cost of effort both to production workers and managers is equal to e, and this cost is not incurred if they shirk. Owners undertake the monitoring of managers.

If an owner employing L production workers and mL managers exerts effort amL, then each manager is caught shirking with probability $p(a)$. We assume that $q(m)$ and $p(a)$ are smooth, increasing, and concave, with $q(0)=p(0)=0$ and $q(.),\ p(.) < 1$. We denote per period wages of production workers by w and the salaries of managers by s. To focus on our main interest, we are making the extreme assumption that managers are not directly productive: their only role is to gather information and monitor production workers.

This model is closely related to the one studied by Shapiro and Stiglitz (1984). The main differences from their analysis are that (i) q is endogenous; (ii) some workers are employed as supervisors; (iii) there is greater scope for contracting. On this last point recall that Shapiro and Stiglitz assume that if a worker is caught shirking, he suffers no monetary penalty but is simply fired. In contrast, in our previous analysis, we assumed a worker only gets his wage if he is not caught shirking. Reality presumably lies somewhere in between. It is difficult to retain all of the past wages, but workers lose their bonuses, their chances of promotion and their pensions when they are fired. We will model this in a simple reduced form way by supposing that a worker (or manager) who is caught shirking can be made to suffer a financial loss of αw (or αs), where $\alpha \geqslant 0$. [14]

3.2. Characterization of equilibrium

Since there are no adjustment costs, every period firm i maximizes

$$\Pi_i = AF(L_i) - w_i L_i - s_i m_i L_i - a_i m_i L_i \tag{8}$$

by choosing L_i, s_i, w_i, m_i, and a_i subject to a participation constraints and incentive compatibility constraints for each occupation.

To write the incentive compatibility constraints, we need to work with Bellman equations. Let us define $V_{\mathrm{U}}, V_{\mathrm{E}}^P, V_{\mathrm{E}}^M, V_{\mathrm{S}}^P, V_{\mathrm{S}}^M$, respectively, as the expected present discounted values of unemployment; employment as a production worker and exerting effort; employment as a supervisor and exerting effort; employment as a production worker and shirking; and employment as a supervisor and shirking. We will use i as an additional argument to indicate when these values are in principle different across firms. Following Shapiro and Stiglitz (1984), we will concentrate on steady states and impose that the time derivatives for all these value functions are equal to zero.

Using standard arguments, we can write:

$$rV_{\mathrm{U}} = z + x[\mu V_{\mathrm{E}}^M + (1-\mu)V_{\mathrm{E}}^P - V_{\mathrm{U}}], \tag{9}$$

where z is utility when unemployed, x is the probability (flow rate) of getting a job and μ is the fraction of jobs that are managerial. Intuitively, an unemployed worker gets a job with probability x; with probability μ, this is a management job, and with

[14] Notice that if the firm could only retain the promised payment for the last instant of work, we would have $\alpha \to 0$. Therefore, for $\alpha > 0$, we need the firm to punish the worker by retaining the payments from an interval of time.

probability $1-\mu$, he becomes a production worker. In both cases he gains the expected present value of the relevant job and loses the present value of unemployment.

All firms take the value of unemployment V^{U} as given by the market, but through their choice of wages and corporate structure affect all other values, hence they are indexed by i. For firm i, we have

$$\begin{aligned} rV_{\mathrm{E}}^{P}(i) &= w_i - e + b[V_{\mathrm{U}} - V_{\mathrm{E}}^{P}(i)], \\ rV_{\mathrm{E}}^{M}(i) &= s_i - e + b[V_{\mathrm{U}} - V_{\mathrm{E}}^{M}(i)], \end{aligned} \tag{10}$$

where w_i is the worker's wage and s_i the manager's salary in firm i, and b is the exogenous flow rate at which jobs dissolve. In equilibrium since all workers exert effort there are no firings for shirking. However, the value of shirking will be important in determining the incentive compatibility conditions. These are written as

$$\begin{aligned} rV_{\mathrm{S}}^{P}(i) &= w_i - q_i\alpha w_i + (b+q_i)[V_{\mathrm{U}} - V_{\mathrm{S}}^{P}(i)], \\ rV_{\mathrm{S}}^{M}(i) &= s_i - p_i\alpha s_i + (b+p_i)[V_{\mathrm{U}} - V_{\mathrm{S}}^{M}(i)]. \end{aligned} \tag{11}$$

The main difference between (10) and (11) is that in (11), there is no cost of effort e, but the relation comes to an end faster as shirking employees are caught (at the rates q and p). When they are caught, they also lose a proportion α of their wages.

The two incentive compatibility constraints are: $V_{\mathrm{E}}^{P}(i) \geqslant V_{\mathrm{S}}^{P}(i)$ and $V_{\mathrm{E}}^{M}(i) \geqslant V_{\mathrm{S}}^{M}(i)$. Simple algebra enables us to write these incentive compatibility constraints as

$$\begin{aligned} w_i &\geqslant \frac{(e/q_i)(r+b) + z + e + rV_{\mathrm{U}}}{(r+b)\alpha + 1}, \\ s_i &\geqslant \frac{(e/p_i)(r+b) + z + e + rV_{\mathrm{U}}}{(r+b)\alpha + 1}. \end{aligned} \tag{12}$$

As in the previous section, both incentive compatibility constraints will bind (otherwise m or a could be reduced).

Since there are two different occupations, there are also two participation constraints:

$$\begin{aligned} V_{\mathrm{E}}^{P}(i) &= \frac{w_i - e + bV_{\mathrm{U}}}{r+b} \geqslant V_{\mathrm{U}}, \\ V_{\mathrm{E}}^{M}(i) &= \frac{s_i - e + bV_{\mathrm{U}}}{r+b} \geqslant V_{\mathrm{U}}. \end{aligned} \tag{13}$$

Note that V_{U}, the value of unemployment, is playing a dual role here. First, it is very similar to $\underline{u}$ in Section 2, the ex ante reservation utility. But here V_{U} also enters into the ex post reservation utility: A worker who is caught shirking receives not 0 as he did in Section 2, but $V_{\mathrm{U}} - \alpha w$. In fact, if $\alpha = 0$, these two concepts will coincide. In contrast, when $\alpha > 0$, the ex ante constraint is always more binding than the ex post constraint. This also gives the basic intuition for why, as long as $\alpha > 0$, our results in this section will be qualitatively similar to those from the static model where the only constraint was the ex ante one.

Now, the problem of firm i is to maximize (8) subject to (12) and (13). This problem still consists of a recursive set of strictly convex optimization problems and therefore the solution is unique. So $w_i = w$, $s_i = s$, $m_i = m$, $a_i = a$ and $L_i = L$ for all i. In particular, we can first determine s and a, then w and q and then finally, L, which will once more simplify the analysis.

A steady state equilibrium is then a vector $(\hat{a}, \hat{s}, \hat{m}, \hat{w}, \hat{L}, \hat{V}_{\mathrm{U}})$ in which (i) the sub-vector $(\hat{a}, \hat{s}, \hat{m}, \hat{w}, \hat{L})$ maximizes (8) subject to (12) and (13) given $\hat{V}_{\mathrm{U}}$; (ii) $\hat{L}(1+\hat{m}) \leqslant N$; and (iii) $\hat{V}_{\mathrm{U}}$ solves (9) with $x = b(1+\hat{m})\hat{L}/(N-(1+\hat{m})\hat{L})$ and $\mu = \hat{m}/(1+\hat{m})$. [15] Intuitively, an equilibrium requires that given the reservation utility of workers, firms choose the optimal wage, salary and organizational forms and then the reservation utility of workers be determined consistently in general equilibrium.

Proposition 4. A steady state equilibrium $(\hat{a}, \hat{s}, \hat{m}, \hat{w}, \hat{L}, \hat{V}_{\mathrm{U}})$ always exists.

The proof employs standard arguments and is sketched in Appendix A. In contrast to Proposition 2 uniqueness is no longer guaranteed because of the general equilibrium interactions determining the value of unemployment, V_{U}.

The equilibrium takes one of three forms, depending on which of the two participation constraints bind:

1. Full employment equilibrium (FEE) where both participation constraints in (13) hold as equality. In this equilibrium $\hat{s} = \hat{w} = r\hat{V}_{\mathrm{U}} + e$, $\hat{L} = N$ and $\hat{m}$ and $\hat{s}$ solve given (12).
2. Unemployment equilibrium (UE) where both participation constraints in (13) are slack and $\hat{L} < N$. In this regime $(\hat{a}, \hat{s}, \hat{m}, \hat{w}, \hat{L})$ maximizes (8) subject to (12) only.
3. Semi-constrained equilibrium (SCE) where one of the participation constraints in (13) hold and the other is slack.

The semi-constrained equilibrium can have either the participation constraint of workers or managers bind, but we think of the case where that of the managers hold, so that $s > w$, as more relevant. The recursive structure of the problem once again helps in the analysis. In the full employment equilibrium, the market dictates what wages must be paid, and thus $\hat{w} = \hat{s} = r\hat{V}_{\mathrm{U}} + e$. Once wages are determined, then the firm minimizes its costs by minimizing monitoring which entails setting $\hat{m}$ and $\hat{s}$ to solve (12). This has an obvious similarity to the full-employment regime of the static model. In contrast, in the unemployment equilibrium, both participation constraints are slack, so the firm is unconstrained by the market and can choose the wage and monitoring levels that maximize profits: [16] $\hat{w} = w^*$, $\hat{s} = s^*$, $\hat{m} = m^*$ and $\hat{a} = a^*$. In other words, as in Section 2, when the ex ante reservation utility, $\hat{V}_{\mathrm{U}}$, is sufficiently low that the firm does not have to compete with other firms to obtain workers, it can attain its "market-unconstrained" optimum. In contrast, in the FEE, $\hat{V}_{\mathrm{U}}$ was sufficiently high that the firm was forced to pay $\hat{s} > s^*$ and $\hat{w} > w^*$ and choose $\hat{q} < q^*$ and $\hat{a} < a^*$.

An important special case is when $\alpha = 0$. In this case, there is no possibility to contract on the wage of the worker when he is caught shirking, so that he receives exactly the

[15] Provided that $\hat{L}(1+\hat{m}) < N$; if $\hat{L}(1+\hat{m}) = N$, (9) becomes $V_{\mathrm{U}} = \mu V_{\mathrm{E}}^{M} + (1-\mu)V_{\mathrm{E}}^{P}$.

[16] That is the firm is maximizing (8) subject to (12) alone.

same payment as when he is not caught shirking. In this case, a full employment equilibrium is not possible. To see this, note that if the participation constraint binds, then $w = rV_U + e$. Substituting this into (12) and setting $\alpha = 0$ gives a contradiction. This is the case considered by Shapiro and Stiglitz (1984), though without endogenous monitoring. As a result, in their model, equilibrium always entails some positive level of unemployment. In contrast, the same exercise shows that when $\alpha > 0$, there will exist a sufficiently high level of V_U such that (12) can be satisfied with $w = rV_U + e$, thus giving a FEE.

3.3. Comparative statics

Let us start the comparative statics with the full employment equilibrium (which, recall, is only possible when $\alpha > 0$). The following proposition is proved in Appendix A:

Proposition 5. In the FEE, $\mathrm{d}m/\mathrm{d}A < 0$, $\mathrm{d}w/\mathrm{d}A > 0$, $\mathrm{d}s/\mathrm{d}A > 0$, $\mathrm{d}a/\mathrm{d}A < 0$.

The intuition is exactly the same as in the static model. In the FEE, an improvement in A increases wages (and salaries) and thus makes workers incentives more *powerful*. This moves firms along both the incentive compatibility constraints of workers and of managers, and both types of employees are monitored less.

Next, let us turn to the unemployment equilibrium. Here, in contrast to the full employment equilibrium, multiple equilibria are possible, and we have to make sure that we are doing comparative statics on the right equilibria. As is well-known in models of multiple equilibria, it is most sensible to look at the extremal equilibria, here defined as those with the highest or lowest value of unemployment, V_U. Then we can state (proof in Appendix A):

Proposition 6. Consider extremal UE. Then $\mathrm{d}m/\mathrm{d}A < 0$, $\mathrm{d}w/\mathrm{d}A > 0$, $\mathrm{d}s/\mathrm{d}A > 0$ *and* $\mathrm{d}a/\mathrm{d}A = 0$.

The intuitive reason for this result is that when A goes up, there is more demand for labor and therefore, wages, and together with wages, salaries increase. One may conjecture that as in the static model, q would remain unchanged because the participation constraints are not binding. However, this conjecture is incorrect due to the cost-of-monitoring effect: The salaries paid to managers are part of the cost of monitoring, and the *cost* of monitoring is higher due to the higher managerial salaries dictated by the market. When monitoring is more costly, firms will want to use less of it; a more buoyant labor market therefore leads to less monitoring and more discretion for production workers.

Similar arguments can also be developed for the case of the semi-constrained equilibrium, and we omit this case. It can be noted at this point that if we were to endogenize monitoring in the exact equivalent of Shapiro and Stiglitz's (1984) set-up with $\alpha = 0$ and no cost-of-monitoring effect, then we would have $\mathrm{d}m/\mathrm{d}A = \mathrm{d}a/\mathrm{d}A = 0$, that is corporate structure would not respond to changes in the state of labor demand. Hence either

$\alpha > 0$ or the cost-of-monitoring effect are necessary for the state of labor demand to affect corporate structure as in our static model.

It is also straightforward to see that, in this dynamic economy, a binding wage floor due labor market regulations or wage setting by unions will work exactly as before. It will push up wages, and therefore induce firms to reduce monitoring. So, the dynamic model also predicts that European economies characterized with more wage push should have less monitoring. We state this as a result and omit the proof:

Proposition 7. Suppose that $\underline{w}$ *is a wage floor imposed by the government. Then, in any steady state equilibrium,* $\mathrm{d}m/\mathrm{d}\underline{w} \leqslant 0$.

3.4. Welfare

Net flow surplus (or net flow output) is: $Y = AF(L) - (1+m)Le - mLa$, where total production is given by the number of production workers, and total effort number of workers in employment is $(1+m)L$ and they incur the effort cost e and finally, owners incur the monitoring cost a for each monitor, thus a total of mLa.

Proposition 8. The decentralized equilibrium never maximizes net surplus.

This proposition again follows by noting that the planner would increase wages and salaries in order to reduce monitoring until there are zero-profits, but in the decentralized equilibrium firms are making positive profits. Taxing profits and subsidizing s and w increases total production as more workers can become producers rather than supervisors. [17]

4. Income distribution

The distribution of income is tied to corporate structure because corporate structure determines both the earnings of production workers, those of managers, and also what fraction of workers become managers. Also, given that cross-country differences in corporate structure appear to be correlated with wage inequality patterns (i.e. the U.S., the U.K. and Canada have experienced sharper increases in wage inequality than other countries in our sample, e.g. Katz et al. (1995)), it is important to investigate the links between the evolution of corporate structure and income distribution. To address this question, we consider a variant of the model in which there are two types of workers.

The two types of workers are capable of doing different kinds of jobs. N "unskilled" workers can only work in production. H "skilled" (college graduate) workers can either work as managers and monitor workers or they can work as engineers. Total output from production workers is equal to $F(L)$ as before; the measure of firms is still 1. Engineering output is given by $\Phi(E)$, where E is the total number of engineers; we

[17] Note that in this case, there are additional issues because the lower of unemployment induces workers to shirk more, thus creating a negative externality on firms. However, as in the original Shapiro and Stiglitz (1984) model, this effect is always dominated.

assume that there are no incentive problems for workers in the engineering sector. As before, monitors are not directly productive. We make the standard assumptions on both production functions: F and Φ are increasing and strictly concave and they satisfy Inada type conditions. We assume that entry into the engineering sector is free and each engineer is paid his marginal product. Also, college graduates do not increase their probability of getting into managerial jobs by being unemployed: they can equally well work as engineers and still receive offers of management jobs.[18]

As in the previous section, the flow rate of detecting a worker who shirks is $q(m)$ where $1/m$ is the number of workers monitored by one manager and the flow rate of detecting a shirking manager is $p(a)$; p and q are both increasing and strictly concave. As before, all agents are risk-neutral, infinitely lived and discount the future at the rate r.

Firm i now maximizes: $AF(L_i) - w_iL_i - s_im_iL_i - a_im_iL_i$ where s_i is salary for the monitors and w_i is the wage rate of the workers. Let us define, $V_{\mathrm{E}}^P, V_{\mathrm{E}}^M, V_{\mathrm{S}}^P, V_{\mathrm{S}}^M$ as the value functions of working and shirking managers and workers. Also differently from the previous section, we need two reservation utilities: V_{U}, the value of unemployment for unskilled workers, and V_{C}, the value of working in the engineering sector for college graduates, which will act as the ex ante and ex post reservation utility for college graduates since they can always choose this option.

Eqs. (10) and (11) determine the value functions as before with the only change that for production workers, the reservation utility is V_{U} and for managers it is V_{C}. Combining these two equations, we can write the incentive compatibility constraints in this case as

$$w_i \geqslant \frac{(e/q(m_i))(r+b)+z+e+bV_{\mathrm{U}}}{(r+b)\alpha+1} \quad \text{and} \quad s_i \geqslant \frac{(e/p(a_i))(r+b)+z+e+bV_{\mathrm{C}}}{(r+b)\alpha+1}.$$

And the two participation constraints are as before: $V_{\mathrm{E}}^P(i) \geqslant V_{\mathrm{U}}$ and $V_{\mathrm{E}}^M(i) \geqslant V_{\mathrm{C}}$. Once more, the maximization problem of firm i subject to the incentive and participation constraints is strictly concave, thus has a unique solution. Therefore, we have $w_i = w$, $s_i = s$, $m_i = m$, $a_i = a$ and $L_i = L$.

The Bellman equation for reservation utility of unskilled workers is

$$rV_{\mathrm{U}} = z^P + x^P[V_{\mathrm{E}}^P - V_{\mathrm{U}}], \tag{15}$$

where z^P is the unemployment benefit for production workers, and x^P is their job-finding rate, which in steady state is equal to $x^P = bL/(N-L)$. The reservation utility of college graduates can be written as follows:

$$rV_{\mathrm{C}} = \Phi'(E) + x^m[V_{\mathrm{E}}^m - V_{\mathrm{C}}], \tag{16}$$

where $\Phi'(E)$ is the wage they receive in the engineering sector and $x^m = bmL/E$ is the rate at which engineers get managerial job offers. Market clearing for college graduates implies: $E = H - mL$.

[18] Thus, there will be no “unemployment” of college graduates; instead there may be equilibria in which engineers would strictly prefer to be managers.

An equilibrium again exists, and takes one of several forms, as before, depending on whether participation constraints are binding or not. The comparative static results are also very similar to those in Section 3. In particular in the full employment equilibrium, an increase in A (labor demand) leads to higher wages and to lower m, that is to less monitoring. What is different, however, is that this increase in A will increase E, the number of college graduates who go into engineering, and thus reduce Φ' and s. Hence, the prediction of the two-type model is that starting from a full employment equilibrium, an increase in the productivity of production workers will increase their wages, while at the same time, *reducing* the extent of monitoring and the salaries of managers.

Next consider the equilibrium in which the participation constraint of workers is slack, but that of managers is binding. Again as in Section 3, focusing on extremal equilibria, an increase in A increases labor demand at given m, and this leads to a larger number of skilled workers employed as managers (i.e. mL increases). As a result, E falls *increasing* s. When s increases, the privately optimal amount of monitoring, m, falls (immediately from the first-order condition of the firm with respect to m). Therefore, in this case, higher productivity of workers leads to higher wages both for managers and workers, and to less monitoring. The contrast between this regime and the full employment regime is interesting. In particular, it implies that a reduction in labor demand will reduce wages of all types of labor in the unemployment equilibrium whereas in the full employment equilibrium, it will reduce production workers' wages but increase managerial wages and thus inequality.

5. Concluding comments

This paper has developed a simple approach to the macroeconomics of organization. In our model, organizational forms are designed to provide incentives to workers. When workers cannot be subjected to arbitrarily severe punishments, low wages naturally imply weak incentives, and firms are induced to choose organizational structures that increase monitoring. Wages may be high either because of labor demand variations or because of labor market regulations. In particular, when labor demand increases, firms reduce monitoring for two reasons: (i) workers are paid higher wages and have better incentives; (ii) monitor's salaries also increase and thus monitoring becomes more expensive. Counteracting these two forces, when labor demand is higher, unemployment is low and does not act as an effective discipline device. Nevertheless we show that this effect is always dominated by (i) and (ii). We argue that these effects may help to explain why organizations differ across countries and over time. The model also shows that the organizational differences can have significant implications for macroeconomic performance.

The notion that how the labor market is organized and the strength of labor demand will affect the internal organization of the firm clearly goes beyond the simple application we considered here. In the working paper version, we discussed how the same principles could affect whether firms use long-term contracts, and similar ideas can be applied to the analysis of whether firms use tournaments, deferred compensation, and bonus pay. We view these as interesting areas for future research.

Acknowledgements

We thank Robin Greenwood for research assistance, Philippe Aghion, Abhijit Banerjee, Peter Diamond, Raquel Fernandez, Bruce Greenwald, Bengt Holmstrom, Boyan Jovanovic, and seminar participants at the NBER Summer Institute, Yale, the Federal Reserve Bank of Cleveland, NYU, Nuffield College, the CEPR Conference on Rising Inequality, Brussels, and University College London for helpful discussion and comments. Acemoglu acknowledges financial support from the NSF.

Appendix A. Proofs of Propositions 4–6

Here we sketch the proofs of the three propositions in Section 3. The remaining results in the text have very similar proofs which are not repeated here.

Proof of Proposition 4. The maximization problem of each firm (8) subject to (12) and (13) for given $V_{\mathrm{U}} \equiv V$ defines: $a(V)$, $s(V)$, $m(V)$, $w(V)$ and $L(V)$. As established in the text, these are all functions; by the maximum theorem, they are continuous in V. It is straightforward to establish that $(1+m(V))L(V)$ is a decreasing function of V, and therefore, the equation $(1+m(V))L(V)=N$ has a unique solution, which we denote by $\bar{V}$. It also follows immediately from the same monotonicity that $(1+m(V))L(V) < N$ if and only if $V > \bar{V}$.

Substituting the two value functions in (10) into the right-hand side of (9) we define, for $V > \bar{V}$,

$$G_1(V) \equiv \frac{1}{r}\,\frac{(r+b)z + x(V)[(m(V)s(V)+w(V))/(1+m(V)) - e]}{(r+b)+x(V)}, \tag{A.1}$$

where

$$x(V) = \frac{b(1+m(V))L(V)}{N-(1+m(V))L(V)}.$$

AS $V \downarrow \bar{V}$,

$$G_1(V) \to \frac{1}{r}\left[\frac{m(V)s(V)+w(V)}{1+m(V)} - e\right] \equiv G_2(V).$$

Therefore an equilibrium, by construction, corresponds to a fixed point $\hat{V}$ of

$$G(V) = \begin{cases} G_1(V), & V > \bar{V}, \\ G_2(V), & V \leqslant \bar{V}, \end{cases} \tag{A.2}$$

provided $L(\hat{V}) \leqslant N$. We will now prove that (A.2) has a fixed-point that satisfies this property.

First observe that $G(V)$ is continuous. Next, we show that $G(V)$ is bounded by showing that both of its components are. To start with, since G_2 is continuous on the compact domain $[0, \bar{V}]$, it is bounded above. Next, write G_1 as

$$\frac{1}{r}\,\frac{(r+b)z(N-(1+m(V))L(V))+b(1+m(V))L(V)[(m(V)s(V)+w(V))/(1+m(V))-e]}{(r+b)(N-(1+m(V))L(V))+b(1+m(V))L(V)}.$$

Because $N > (1+m(V))L(V) \geqslant 0$ on $V \in (\bar{V}, \infty)$, the denominator is bounded by bNr and $(r+b)Nr$; the numerator is bounded below by 0. Also because

$$(1+m(V))L(V)\left[\frac{m(V)s(V)+w(V)}{1+m(V)}-e\right]$$
$$< L(V)[m(V)s(V)+w(V)] < AF(L(V)) < AF(N)$$

(the second inequality because maximized profit is always nonnegative), the numerator is bounded above by $(r+b)zN + bAF(N)$. Thus G_1 is bounded above and below, proving that $G(V)$ is a bounded function.

Now consider the continuous, bounded function

$$H(V) = G(V) + \max\left\{\min\left\{1, \frac{(1+m(V))L(V)}{N}-1\right\},\ 0\right\}.$$

Since H continuously maps a compact domain onto itself, it has a fixed point $\hat{V}$. We claim that this is an equilibrium.

If $(1+m(\hat{V}))L(\hat{V}) \leq N$, $\hat{V}$ is also a fixed point of G and is an equilibrium by construction. To complete the proof we need to show that $(1+m(\hat{V}))L(\hat{V}) \leqslant N$. Suppose $(1+m(\hat{V}))L(\hat{V}) > N$: then $\hat{V} < \bar{V}$ and so

$$\hat{V} = H(\hat{V}) > G_2(\hat{V}) = \frac{1}{r}\left[\frac{m(\hat{V})s(\hat{V})+w(\hat{V})}{1+m(\hat{V})}-e\right].$$

But adding the two participation constraints (13) together gives

$$\hat{V} \leqslant \frac{1}{r}\left[\frac{m(\hat{V})s(\hat{V})+w(\hat{V})}{1+m(\hat{V})}-e\right],$$

which is a contradiction. This establishes the claim. □

Proof of Proposition 5. Let $r\hat{V}_{\mathrm{U}} + e = v$. For full employment we have that

$$AF'\left(\frac{N}{1+m(v)}\right) = v + m(v)(v + a(v)), \tag{A.3}$$

where $a(v)$ and $m(v)$ solve (12). $N/(1+m(v))$ is the number of production workers that need to be employed when the monitoring level is given by $m(v)$ in order to ensure full-employment. It is straightforward to see that, since p and q are concave, $a(v)$ and $m(v)$ are decreasing functions of v.

Next note that the firm is actually choosing s and a subject to the constraint that $s \geqslant v$, thus $s = v$ if and only if $\partial\Pi(s=v)/\partial s < 0$. Thus, we have $1 + a'(v) > 0$. By the same argument regarding the choice of w and m, we have $1 + m'(v)(v + a) > 0$. Therefore, the right-hand side of (A.3) is increasing in v. In contrast, the left-hand side is decreasing in v, since $m'(v) < 0$ and $F'' < 0$. Therefore, a full employment equilibrium, when it exists, is uniquely defined. Now an increase in A raises the left-hand side, thus requires an increase in the right-hand side, hence an increase in v. $\mathrm{d}w/\mathrm{d}A > 0$, $\mathrm{d}s/\mathrm{d}A > 0$, $\mathrm{d}m/\mathrm{d}A < 0$ and $\mathrm{d}a/\mathrm{d}A$ immediately follow from $\mathrm{d}v/\mathrm{d}A > 0$. □

Proof of Proposition 6. UE is characterized by:

$$\begin{aligned}
&AF'(L^*) - w^* - m^*(s^* + a^*) = 0,\\
&-\frac{(r+b)}{(r+b)\alpha+1}\,\frac{e}{Q(m^*)^2}Q'(m^*) + s^* = 0,\\
&-\frac{(r+b)}{(r+b)\alpha+1}\,\frac{e}{P(a^*)^2}P'(a^*) + 1 = 0,
\end{aligned} \tag{A.4}$$

and also (12) and (9). It is then straightforward to see that a^* is fixed, but all other variables vary with $V_{\mathrm{U}} \equiv V$, thus we have $s^*(V)$, $m^*(V)$ and $w^*(V)$ with s^* and w^* as increasing functions of V and m^* as a decreasing function of V. Then, substituting into (9), we obtain

$$V = G(V) \equiv \frac{(r+b)z + x(V)[(m^*(V)s^*(V) + w^*(V))/(1+m^*(V)) - e]}{(r+b)r + bx(V)}, \tag{A.5}$$

where

$$x(V) = \frac{b(1+m^*(V))L^*(V)}{N - (1+m^*(V))L^*(V)}.$$

Since the right-hand side of (A.5), $G(V)$, is a non-linear function, we cannot establish uniqueness of unemployment equilibrium. But it is clear that $G(0) > 0$ and also $\lim_{V\to\infty} G(V) < \infty$. Thus, the extremal equilibria always have $G(V)$ cutting the 45° line from above. Next, note that an increase in A for given V only affects L^*, thus x. In particular, x increases when A goes up. Hence, a higher A shifts $G(V)$ up, therefore, at extremal equilibria: $\mathrm{d}V/\mathrm{d}A > 0$. This immediately implies that at extremal equilibria, $\mathrm{d}w/\mathrm{d}A, \mathrm{d}s/\mathrm{d}A > 0$ and $\mathrm{d}s/\mathrm{d}A < 0$, but a remains at a^*. □

References

Acemoglu, D., 1999. Changes in unemployment and wage inequality: An alternative theory and some evidence. American Economic Review 89, 1259–1278.

Appelbaum, E., Batt, R., 1994. The New American Workplace. ILR Press, New York.

Bulow, J., Summers, L., 1986. A theory of dual labor markets with application to industrial policy, discrimination and keynesian unemployment. Journal of Labor Economics 4, 376–415.

Calvo, G., Wellisz, S., 1979. Hierarchy, ability and income distribution. Journal of Political Economy 87, 991–1010.

Caroli, E., van Reenen, J., 1999. Wage inequality and organizational change. Mimeo., University College London.

Chandler, A.E., 1977. The Visible Hand: The Managerial Revolution in American Business. Harvard University Press, Cambridge, MA.

Foster, J.E., Wan, H.Y., 1984. Involuntary underemployment as a principal–agent equilibrium. American Economic Review 74, 476–484.

Freeman, R., 1995. Are your wages set in Beijing? Journal of Economic Perspectives 9, 15–32.

Freeman, R., Lazear, E., 1994. An economic analysis of works councils. In: Rogers, J., Streeck, W. (Eds.), Works Councils: Consultation, Representation and Cooperation in Industrial Relations. University of Chicago Press, Chicago.

Gibbons, R., 1996. Incentives and careers in organizations. In: Kreps, D., Wallis, K. (Eds.), Advances in Economics and Econometrics: Theory and Applications. Cambridge University Press, Cambridge.

Gordon, D., 1996. Fat and Mean: The Corporate Squeeze of Working Americans and the Myth of Corporate Downsizing. Martin Kessler Books, New York.

Holmstrom, B., 1979. Moral hazard and observability. Bell Journal of Economics 13, 324–340.

Katz, L., 1987. Efficiency wage theories: A partial evaluation. In: Blanchard, O., Fischer, A.S. (Eds.), NBER Macroeconomic Annual 1986, MIT Press, Cambridge, MA, pp. 176–235.

Katz, L., Loveman, G.W., Blanchflower, D.G., 1995. A comparison of changes in the structure of wages in four OECD countries. In: Freeman, R., Katz, L. (Eds.), Differences and Changes in Wage Structures. National Bureau of Economic Research, University of Chicago Press, Chicago.

Kremer, M., Maskin, E., 1998. Wage Inequality and Segregation by Skill. National Bureau of Economic Research Working Paper no. 5718.

Legros, P., Newman, A.F., 1996. Wealth effects, distribution and the theory of organization, Journal of Economic Theory 70 (2), 312–341.

Prendergast, C., 1999. The provision of incentives and firms. Journal of Economic Literature 28, 7–63.

Rebitzer, J.B., Taylor, L.J., 1991. A model of dual labor markets when product demand is uncertain. Quarterly Journal of Economics 106 (4), 1373–1383.

Shapiro, C., Stiglitz, J., 1984. Unemployment as a worker discipline device. American Economic Review 74, 433–444.

Weiss, A., 1990. Efficiency Wages: Models of Unemployment, Layoffs and Wage Dispersion. Princeton University Press, Princeton, NJ.