DIVISION OF THE HUMANITIES AND SOCIAL SCIENCES

CALIFORNIA INSTITUTE OF TECHNOLOGY

PASADENA, CALIFORNIA 91125

# ORGANIZATIONAL DISECONOMIES OF SCALE

R. Preston McAfee
California Institute of Technology
University of Western Ontario

John McMillan
University of California, San Diego

SOCIAL SCIENCE WORKING PAPER 728

April 1990

# ORGANIZATIONAL DISECONOMIES OF SCALE*

**R. Preston McAfee and John McMillan**

California Institute of Technology and University of Western Ontario; and University of California, San Diego

March 1990

*Abstract*

This paper models strategic behavior within firms. The principal (e.g., the firm's owner) is handicapped by not knowing as much about the firm's capabilities as the agent(s) (e.g., the manager). The agent can extract some rents from his private information. The principal can retrieve some of these rents at the expense of introducing a distortion, paying the agent less than the full value of his marginal product. As a result the firm operates inefficiently. The degree of this inefficiency varies with demand elasticity and with the length of the firm's managerial hierarchy. The costs of operating the hierarchy create a limit to the size of the firm.

Correspondence to:

Preston McAfee
Division of Humanities and Social Sciences
California Institute of Technology
Pasadena, CA 91125, USA

## I. Introduction

A spate of recent articles in the business press suggest that small is becoming beautiful. With titles like "Is Your Company Too Big?", "The Age of Hierarchy Is Over", "Forget Hierarchy", and "The Battle of the Bulge", they report that large firms are reducing the number of layers of management; becoming smaller by focusing on a narrower range of activities and contracting out the rest; and mimicking small firms by creating decentralized subunits with responsibility for their own decision-making and profits.[1]

What are the limitations to the size of firms? Why can a firm not achieve at least constant returns to scale simply by replicating the basic production process? Why, to quote Williamson (1985, p.131), "can't a large firm do everything that a collection of small firms can do and more?" It seems clear that the answer lies in the difficulties of controlling a large organization. It is less clear how to make this answer precise.

Organizational costs are multifaceted; in this paper we look at one source of organizational diseconomies. We model a firm as a hierarchical structure, focusing on the possibilities for strategic behavior that arise as a result of the information about the firm's capabilities being dispersed within the firm. The simplest version of our model has a principal designing the incentives within which an agent works. Principal and agent can be interpreted as owner and manager; or top management and divisional management; or supervisor and worker. The agent has some information about the firm's productivity that the principal does not have: detailed knowledge of how the firm's production processes work, obtained from taking part in the day-to-day production activities (in the phrase of Hayek (1945), "knowledge of the particular circumstances of time and place"). This knowledge gives the agent bargaining power: he can extract some rents from his private information. The principal can retrieve some of these rents at the expense of introducing a distortion, paying the agent less than the



full value of his marginal product. As a result the firm operates inefficiently, with the degree of inefficiency determined by model's parameters.

We next model a multi-tier hierarchy. We suppose that the workers at the bottom of the hierarchy have private information. As a result of the individuals' exploitation of their informational advantages, operating the hierarchy gives rise to information costs, which increase surprisingly quickly as the hierarchy lengthens: in one example, the informational losses double whenever an extra layer is added to the hierarchy.

Though simple and standard, our model is rich in predictions:

-- With production efficiency measured as the ratio of lowest possible cost to cost actually incurred, a firm's efficiency falls as its hierarchy lengthens.

-- When the firm's market power falls (i.e., the elasticity of demand for its output rises) production efficiency rises or falls depending on the form of the cost function.

-- A firm may respond to adversity by becoming more efficient.

-- The longer the hierarchy, the smaller the marginal rate of payment with respect to output of the workers at the bottom of the hierarchy. (Thus small firms will tend to pay their workers piece rates, and large firms will pay fixed wages.)

-- The higher an individual is up a hierarchy, the more sensitive are his marginal payments to performance. (Thus bonuses will be a bigger fraction of income for executives than for production-line workers.)

-- The more competitive the firm's output market, the more sensitive pay is to performance. (So competitive firms will tend to pay their workers piece rates, and monopolists will pay close to fixed wages.)

-- A firm with a long hierarchy may not be viable in a competitive industry. (So a large firm exposed to competition by the opening of international trade or the introduction of antitrust laws might respond by reducing the number of levels in its hierarchy.)

We conduct the Williamsonian experiment of merging all of the firms in an industry to become the divisions of a monopoly directed by an overall principal; and we compare the performance of the separate firms with the performance of the divisions of the monopoly. We find that the comparison involves a three-way trade-off. There is the standard gain from monopolization: total profits rise because the monopoly can squeeze extra profits from its customers. And there are two contradictory technical-efficiency effects. First, because the monopolized firm has an extra layer of hierarchy, the divisions are optimally given weaker incentives to exert effort than they had as independent firms. They therefore produce further from full-information efficiency. Second, because of the variation in production costs among the firms/divisions, in the equilibrium with competing firms the high-cost firms produce relatively too much and the low-cost firms relatively too little. With monopolization, the owner of the firm can correct this inefficiency, but the information costs mean that he cannot fully correct it. We find that the extra information costs within the merged firm mean it is not being profitable to monopolize the industry if final demand is elastic enough. Thus the model gives a demand-side determinant of the extent of an industry's concentration.

As noted, we focus on one particular source of limits to hierarchy: private information. Our model therefore explores one particular aspect of the costs of centralization that Milgrom (1988) and Milgrom and Roberts (1988, 1989) term "influence costs". In our model, influence costs arise because the person in authority, knowing less than his subordinates, is tempted to make inefficient decisions to limit his subordinates' bargaining advantages. Limits to hierarchy alternative to those modeled here include the limited information-processing capacities of managers (Geanakoplos and Milgrom, 1985, Guesnerie and Oddou, 1988; Williamson, 1967), and the cost of monitoring subordinates and the resulting inadequate effort levels (Calvo and Wellisz, 1978, Mirrlees, 1976, Rosen, 1982).[2]



## 2. Hierarchies with Incomplete Information

Consider first a simple hierarchy consisting of a single agent and a single principal. Both are risk neutral. The agent has a type (e.g., inherent productivity), denoted $t$. The principal perceives the agent's type as being drawn from a distribution $F(t)$, with density $f(t)$ and support [0, 1]. After accepting his contract, an agent chooses an output to produce, $q$. Let $C^0(q, t)$ be the cost to an agent of obtaining a given output $q$ when his type is $t$. We assume that higher types have a lower cost and a lower marginal cost: $C^0_t(q, t) \leq 0$, $C^0_{qt}(q, t) \leq 0$ (where subscripts denote partial derivatives). We assume also that the marginal cost of output is positive and nondecreasing: $C^0_q(q, t) > 0$, $C^0_{qq}(q, t) \geq 0$. The total revenue from selling the output is $R(q)$, with $R''(q) \leq 0$.

Laffont and Tirole (1986) show that the agent extracts rents from his private information, so the cost borne by the principal is higher than the actual production cost. As a result, the principal orders an output from the agent that is less than the efficient level. For the sake of completeness we give a quick derivation of this here (although the reader may skip to equation (4)).

The principal designs the way the agent is remunerated, offering to pay an amount that depends not only on the output the agent produces but also on the type the agent reports himself to be (which we denote by $r$). The agent's expected profit is

$$\text{(1)} \qquad \pi(t) = \max_r E[\text{payment} \mid r] - EC^0(q(r), t)].$$

The Revelation Principle (Myerson, 1982) says that, without loss of generality, we can restrict attention to payment functions that induce the agent to report his private information correctly. Imposing the requirement that truth-telling is optimal for the agent on (1) and using the Envelope Theorem with derivatives evaluated at truth-telling (i.e., $r=t$), we get:

(2) $$\frac{d\pi}{dt} = \frac{\partial\pi}{\partial t} = -EC^0_t(q(t), t).$$

Using (2), integration by parts, and the individual-rationality constraint ($\pi(t)\geq 0$), we can express the agent's expected profit as

(3) $$\int_0^1 \pi(t)f(t)dt = -\int_0^1 h(t)C^0_t(q(t), t)f(t)dt,$$

where $h(t) = [1-F(t)]/f(t)$. Thus the agent earns an informational rent from his private information about his productivity.

The cost borne by the principal in implementing the output $q(t)$ is the actual production cost incurred by the agent plus the agent's informational rent. This cost is, in expectation over $t$, $C^1(q(t), t) = C^0(q(t), t)+\pi(t)$; or, from (3),

(4) $$C^1(q(t), t) = C^0(q(t), t) - h(t)C^0_t(q(t), t).$$

Of course, the output function the principal imposes depends not only on its cost to the principal (4), but also on the principal's benefits from the output, $R(q)$.

An output function $q(t)$ is implementable if and only if $\partial^2\pi/\partial t\partial r$, evaluated at $r=t$, is nonnegative for all $t$; that is, for a slightly higher type, profit is increasing in report. From (1), this condition is equivalent to $EC^0_t(q(r), t)$ being nondecreasing in $r$ at $r=t$, for all $t$. In turn, given $C^0_{qt}\leq 0$, this is equivalent to $dq(t)/dt\geq 0$. Thus an output function is implementable if and only if it asks higher types for more output (McAfee and McMillan, 1987; Rogerson, 1897).

The output the principal wants maximizes $R(q(t))-C^1(q(t), t)$, where $R$ is the total-revenue function. Thus the optimal output satisfies $R'(q(t))-C^1_q(q(t), t) = 0$. Totally differentiating this expression with respect to $t$, we get $dq(t)/dt = C^1_{qt}/[R''-$



$C^1_{qq}]$. Thus, with the assumed concavity of $R$, a sufficient condition for implementability is $C^1_{qt}\leq 0$ and $C^1_{qq}\geq 0$; that is, the principal's cost function $C^1$ inherits the curvature properties of the actual cost function $C^0$. (Unfortunately, since $C^1$ depends on $C^0_t$, the signs of $C^1_{qt}$ and $C^1_{qq}$ depend on third derivatives of $C^0$.)

The second term on the right of equation (4) is the agent's informational rent (which, since $C^0_t$ is negative, adds to the production cost $C^0$). This shows how much the agent must be offered to induce him not to act as though his type is lower than it really is. The principal, because he bears this informational rent, directs the agent to produce less output than the efficient level. In other words, equation (4) says that the principal's asymmetric-information decision problem corresponds to a full-information problem with higher costs (higher by the amount of the information cost).

The informational rent shown by (4) is larger the larger $[1-F(t)]/f(t)$ is; that is, (roughly) the bigger the variation in types across agents. The larger this is, the more bargaining power the agent gets from his private information. The informational rent is also larger the larger $C^0_t$ is; that is, the larger is the effect of type on production cost. To understand the role of $C^0_t$, notice that the cost function $C^0(q, t)$ is a reduced form; the corresponding structural form would explicitly involve the agent's effort. Consider one such structural form. Denote the agent's effort by $y$, and suppose his cost of effort is $\phi(y)$. Define a function $Y(q, t)$ to be the minimal effort needed by an agent of type $t$ to produce the expected output $q$. Then $C(q, t) = \phi(Y(q, t))$, so $C_t = \phi' Y_t$. Here $\phi'$ is the marginal cost of effort, and $Y_t$ is the rate at which the effort necessary to produce $q$ falls as the agent's type rises. Thus $-C_t$ measures the reduced cost to the agent of producing an output appropriate to a lower type; $-C_t$ is the marginal benefit to the agent of imitating a slightly lower type. Thus the larger $-C_t$ is, the more rents the agent to induce earns from his information.

Evidence on how big these rents can be comes from a study of three divisions of large U.S. corporations by Schiff and Lewin (1970). Division managers built slack into

their annual divisional budgets by understating revenues and overstating costs -- inflating personnel requirements, proposing unneeded projects, and failing to report the adoption of cost-lowering process improvements. This slack amounted to an estimated 20% to 25% of the division's budgeted operating expenses.[3] In terms of our model, if we interpret the division as the agent and the top management as the principal, this implies that the agent's rent, $-h(t)C^0_t(q, t)$, is one-fifth to one-quarter of the principal's cost, $C^1$.

Equation (4) embodies the main idea in this paper; the rest of the paper examines the implications of this equation for the costs of running an organization.

## 3. The Efficiency of Production

Do monopolists produce above minimum cost, causing a welfare loss beyond the thoroughly explored allocative inefficiencies? Conversely, does competition force minimum-cost production? Generations of economists have believed that competition provides discipline. The separation between a firm's ownership and its control tends to free the manager to pursue his own aims; but competition from other firms counteracts this by inducing the manager to make relatively efficient production decisions.

A remarkably diverse group of economists agree that monopolies permit inefficiencies. Adam Smith said that monopoly is "a great enemy to good management, which can never be universally established but in consequence of that free and universal competition which forces everybody to have recourse to it for the sake of self-defence" (Smith, 1776, p.165). Hicks (1935, p.8) puts it more pithily: "The best of all monopoly profits is a quiet life." Galbraith (1979, Ch.10), Holmström and Tirole (1987, Section 4.3), Leibenstein (1966) Machlup (1967), Samuelson (1976, pp.508-12), and Scherer (1980, pp.38-41, 464-66) also argue that competitive firms produce more efficiently than monopolies.

Some scattered evidence on this effect exists, summarized by Scherer (1980,



pp.464-66). For example, the breaking up of one U.S. cartel initiated cost-reduction efforts by the member firms that resulted in a decline in manufacturing costs of about one-quarter. As a result of the increased competition that followed the strengthening of Britain's antitrust laws in the 1950s, costs in some industries fell. And in a sample of regulated firms, the presence of product-market competition has been found to be associated with significantly lower unit costs.

Despite the familiarity and plausibility of the idea that competition promotes efficiency in production, it still lacks a convincing theoretical basis.[4] As Holmström and Tirole (1987) note in their survey of the theory of the firm, "Apparently, the simple idea that product market competition reduces slack is not as easy to formalize as one might think".

This "simple idea" has been challenged, suggesting that it is perhaps not so simple. Jensen and Meckling (1976) argue that owners of monopolies have as strong an incentive as owners of competitive firms to prevent any self-seeking behavior by their managers. Thus monopolies should be no less efficient than competitive firms. Equation (4) above shows, contrary to Jensen and Meckling, that when the manager has private information about the firm's capabilities it is not in the owner's interest to induce efficient production; and, as we shall see, the degree of inefficiency that is optimal from the owner's point of view varies depending on whether the firm is monopolistic or competitive.

A natural measure of the efficiency with which a firm produces a given output is the ratio of the lowest possible cost of producing that output to the actual cost. (This is the technical-efficiency measure proposed by Farrell, 1957.) From last section's analysis efficiency, so measured, equals $C^0/[C^0-h(t)C^0_t]$, or $1/[1+r]$, where $r$ is the ratio of information cost to production cost (i.e., $r = -h(t)C^0_t(q, t)/C^0(q, t)$). Thus the extent of inefficiency depends on the size of the information cost relative to production cost. As information costs decline relative to production costs, measured efficiency

increases towards one.

The firm's external environment affects efficiency. The principal decides the quantity he wants the agent to produce by equating marginal revenue to the marginal cost the principal pays, $C^1_q$. Assume the principal faces increasing marginal costs, so that $C^1_{qq} > 0$. ($C^1_{qq}$ equals $C^0_{qq}-\{[1-F(t)]/f(t)\}C^0_{tqq}$, so this requires that $C^0_{tqq}$ is, if positive, not too large.) Consider the experiment of a small rotation of the demand curve about a pre-existing optimal point. This increases the demand elasticity and increases marginal revenue at that point. This means (given the assumed concavity of the total-revenue function) that the principal wants more produced than before. What is the effect of the increased production on efficiency? Efficiency, measured as $1/[1+r]$, rises or falls with increases in output as $r$, the ratio of information cost to production cost, falls or rises. Thus efficiency rises with increases with output if and only if $qC^0_q/C^0$, the elasticity of cost with respect to output, exceeds $qC^0_{tq}/C^0_t$, the elasticity of the rate of change of cost with respect to type. For example, with the cost function $C^0=[z+1-t]q^2$, these elasticities are equal and efficiency is independent of output. With $C^0=zq+(1-t)q^2$, $C^0_t$ is more elastic than $C^0$, and efficiency declines as output increases.

This ambiguity in responses is surprising. As noted, it is commonly asserted that competitive firms produce more efficiently than monopolies. But we have just seen that, depending on the cost-elasticity condition, efficiency may either rise or fall as as the demand facing the firm becomes more elastic, that is, more "competitive". The standard conclusion can be regained in our model if we assume that there are many potential owner-managed firms that could enter the industry and compete with our hierarchical firm. If the elasticity of $C^0$ exceeds the elasticity of $C^0_t$, then the hierarchical firm might be able to produce efficiently enough to survive the competition from owner-managed firms. But if the elasticity of $C^0$ is less than the elasticity of $C^0_t$, the hierarchical firm will be forced to leave the industry, and the industry will consist



only of small firms. Hence the cost-elasticity condition determines the size distribution of firms in a free-entry industry.

If we define profit as in elementary textbooks to be revenue minus production cost, then in the model developed here firms do not maximize profit; rather, they maximize revenue minus cost inflated by the information-cost term. Samuelson (1976, p.508), in assessing the assumption of profit maximization, echoes Hicks on the monopolist's quiet life: "As soon as the firm becomes of any considerable size and begins to enjoy some control over price, *it can often afford to relax a little* in its maximizing activities." According to Samuelson, firms with less elastic demand operate less efficiently. We have seen that this is true in our model if the cost elasticity condition holds; that is, if $C^0_t$ is less sensitive to output variations than is cost itself. But efficiency falls with changes in demand elasticity not because the firm relaxes in its maximizing activities; rather, it is because the information constraints facing the firm's principal change with the firm's environment.

For a similar comparative-statics exercise, suppose there is some parameter $z$ that enters the cost function, with $C^0_z>0$. Then an increase in $z$ increases efficiency if and only if the elasticity of cost $C^0$ with respect to $z$ exceeds the elasticity of $C^0_t$ with respect to $z$. In particular, if the parameter $z$ does not alter the rate at which changes in type affect cost, so $C^0_{tz}=0$, then an increase in $z$ increases efficiency. Newspaper reports often claim that firms respond to adversity by trying harder. For example, it is asserted that Japanese firms responded to the recent rise in the value of the yen by becoming more efficient. In terms of standard microeconomics, this claim is puzzling, for it implies they must not have been optimizing before the yen change. But it is explicable by our model. Denote the exchange rate by $z$. The rise in the yen increased the price of some inputs, so $C^0_z>0$. But plausibly it did not affect the magnitude of the private information within a firm, $C^0_t$, so $C^0_{tz}=0$. Hence the model predicts an increase in the efficiency of production following this external shock.

What contract does the principal offer the agent? The principal wants an output $q^I(t)$ that maximizes his profit, $R(q)-C^I(q, t)$. This output can be evoked by offering the agent a payment that is a linear function of output, with a marginal remuneration rate of $C^0_q(q^I(t), t)$, provided $dC^0_q(q^I(t), t)/dt \geq 0$ (McAfee and McMillan, 1987). This can be interpreted as a piece rate, commission rate, or managerial incentive scheme. That the desired output is less than the efficient level implies that this marginal payment rate is less than 100%. This payment scheme varies with the firm's demand. Consider again the experiment of rotating the demand curve about the existing optimum. As demand becomes more elastic, the desired quantity $q^I(t)$ rises. Because the agent's marginal costs are increasing, this means that his marginal payment rate increases. Thus the more competitive the firm's output market, the more stringent are the contractual incentives offered to the agent.[5]

## 4. Monopoly vs. Competition

Our model gives one answer to Williamson's question about the limitations to the size of firms. Holmström and Tirole (1989) paraphrase Williamson's question as: "why one couldn't repeatedly merge two firms into one and by *selective intervention* accomplish more in the integrated case than in the decentralized case. In other words, let the two firms continue as before and interfere (from the top) only when it is obviously profitable. The fact that there are limits to firm size must imply that selective intervention is not always feasible."

Suppose that, in the Holmström-Tirole merged firm, selective intervention from the top requires that an overall principal be added to control the merged firm. Suppose also that it is feasible for the overall principal to intervene selectively and create efficiency gains. Then what were the two independent firms, and are now divisions of the merged firm, can perform better than before. But, by our argument above, their



costs of operation, as perceived by the new overall principal, incorporate a new informational rent associated with the extra level of hierarchy. The total rents earned within the two divisions/firms may be higher than before the merger; but the overall principal may not be able to appropriate enough of these rents to cover the information costs he must pay. Although there are potential efficiency gains from the merger, it may not pay anyone to organize it.

To compare market coordination with coordination within a firm, let us examine more carefully an experiment of the sort proposed by Williamson and Holmström and Tirole. Imagine an industry consisting of $n$ separate, identical firms. Each of these firms consists of a single entrepreneur/worker. (This can be interpreted as a reduced-form representation of a hierarchical structure.) Each firm has private information about its own type, which determines its costs of production; and it perceives its rivals' types as being independent draws from a distribution $F$. The firms meet each other in the product market in asymmetric-information Cournot quantity competition.

We shall compare this market with the situation after the $n$ firms have been merged and now form the $n$ divisions of a monopolistic firm. Suppose that the problems of bargaining with private information among the $n$ firms (Mailath and Postlewaite, 1988) mean that the $n$ independent firms could not simply form a partnership. Instead, the merged firm must be controlled by a single overall principal, adding a level of hierarchy. A three-way trade-off determines how the merged firm performs in comparison with the independent firms.

Two effects work to produce gains from selective intervention. One is the standard monopoly effect: by choosing the total quantity to be supplied, the monopolist is able to extract more rents from the buyers of the industry's output than are the independent firms.

The second effect can be labeled the rationalization gain from monopolization. It is well known that Cournot competition creates a technical inefficiency when the firms'

costs differ: firms with relatively high cost produce too much output, and firms with low costs produce too little. This inefficiency of competition can be ameliorated in the merged firm: the principal of the merged firm can direct the low-cost divisions to produce relatively more and the high-cost divisions relatively less. But the inefficiencies of hierarchy mean that rationalization gains can only partially be achieved. Assume that the marginal production costs of the firms/divisions increase with output. Full technical efficiency requires that outputs be allocated so that marginal production costs be equated across the divisions. But recall from Section 2 that the costs that the principal bears are not just production costs. Rather, the principal in effect bears production cost plus information cost, so that what he equates across the divisions are these marginal augmented costs. Except in the measure-zero case in which marginal information costs are the same for all divisions, the principal does not induce an efficient allocation of production to the divisions. (Consider the special case in which $F$ is uniform on [0, 1] and the cost function is $C^0(q, t) = (z+1-t)c(q)$ for $z>1$. The marginal cost perceived by the principal (from (4)) is $(z+2(1-t))c'(q)$, and equating this across different divisions with different $t$'s does not in general equate marginal production costs, $(z+1-t)c'(q)$, so there is an inefficiency caused by the adverse selection. Only in the knife-edge case of $z=0$ are marginal production costs equalized among the divisions of the monopoly.)

Working in the opposite direction to these two effects is the information-cost effect derived in Section 2, which tends to make the monopoly produce less efficiently than the independent industry. The independent firms fully bear the costs and benefits of their effort choices; but in the merged firm, the divisions retain only a fraction of the marginal returns from their actions. This inefficiency of hierarchy tends to make the competing firms more profitable than the monopoly.[6]

To examine more carefully the Williamson experiment, comparing a monopolized industry with a competitive industry, consider the following example. There are $n$



producing units, which we shall alternately view as independent firms and divisions of a monopoly. The monopoly is controlled by a single principal; the monopolized industry has one more layer of hierarchy than the competitive industry. The cost function is $zq+(1-t)q^2$, where $t$ is distributed uniformly on [0, 1]. The demand curve is linear: price is $a-bQ$, where $Q=\Sigma_{i=1}^{n} q_i$. Let $\pi^c$ represent expected industry profits (averaged over types) when the firms compete as independent entities; $\pi^m$ total profit (to principal and agents) when the industry is monopolized; and $\pi^l$ the profit earned by the principal when the industry is monopolized. (If $\pi^m$ exceeds $\pi^c$, then it is feasible for the principal to organize a takeover of the $n$ independent firms, paying them their stock-market value $\pi^c$, and still earning positive profit.) Algebraic details are given in the appendix, but the conclusion is that whether or not monopoly does better than competition depends on the parameters. In particular: (1) for $n$ large, $\pi^m \geq \pi^l > \pi^c$; (2) for $b$ large, $\pi^m \geq \pi^l > \pi^c$; (3) for $b$ small, $\pi^c > \pi^m \geq \pi^l$.

The first of these is easily explained. When there are many firms competing, the standard profit increase from monopolization outweighs the organizational costs. Results (2) and (3), showing the effect of the demand curve's slope on the organization of the industry, are more novel. With very elastic demand, two of our three effects disappear: there is no Cournot inefficiency, and there are no profit gains from monopolizing the industry. All that remains are the hierarchical losses due to the asymmetric information. Thus, when industry demand is very elastic, the industry will not be monopolized. This is novel, because it gives a demand-side rationale for industry concentration, complementary to the conventional supply-side argument based on economies of scale. The industry will tend to be competitive if there exist close substitutes for its output (for example, from imports); conversely, industries without close substitutes will tend to be monopolized.

Rents, therefore, are the lubricants that make it possible for a hierarchy to function. This suggests a reversal in the conventional causality of perfectly competitive

industries. We usually think of an industry as being competitive because the firms in it are small. But if larger firms mean longer hierarchies, then rents must be present for a large firm to be viable. Thus firms are small because the industry is competitive.

## 5. Multi-level Hierarchies

Imagine an organization with a pyramidal structure: a single principal oversees a certain number of subprincipals, each of whom supervise some sub-subprincipals, and so on down to the agents at the bottom of the hierarchy. The agents do the actual production, each delivering his output to his immediate supervisor, who in turn delivers it to his immediate supervisor, and so on until the output reaches the overall principal.[7] We do not model why the hierarchical structure is necessary, and why the overall principal does not himself directly supervise the agents' productive activities. Instead, we appeal to the plausible but ad hoc notion that supervision takes time and the principal's time is limited. Thus, once an organization reaches a large enough size, employing many agents, it is not feasible for the overall principal to supervise the agents directly, and he must insert subprincipals between himself and the agents.

Assume that each supervisor designs and administers contracts for the people immediately below him in the hierarchy, who impose the contracts immediately below them. As will be seen, for the hierarchy to be nontrivial in our model, it must not be possible for the overall principal to impose the contracts that apply all the way down the hierarchy: if this were possible, the intermediate principals would be inessential, and it would be as if the overall principal controlled the agents directly. Once again, we appeal to exogenous limits to what it is feasible for the overall principal: he does not have time to administer lower-tier contracts.[8]

We suppose that the agents at the bottom of the hierarchy who produce the output have private information about their types; the different agents draw their types



independently from $F$. Thus an agent's immediate supervisor in designing the contract faces the asymmetric-information problems analyzed in Section 2 above. There is no private information further up the hierarchy: anything an intermediate supervisor does to transform the output is observable by his immediate principal, so from a modeling point of view it is as if the intermediate supervisors simply pass the output up the chain. (This can be regarded as a polar case: if there were additional private information within the hierarchy, the costs of operating the hierarchy that we shall identify would be exacerbated. See Melamud, Mookherjee, and Reichelstein (1989).)

The information cost identified in Section 2 is magnified if the hierarchy has more than one level. Consider a hierarchy with $n$ agents at the bottom, with types $t=(t_1, \ldots, t_n)$, and $k$ levels of principals above them. The top principal cannot prevent the lower-tier principals from contracting with their subordinates and learning the subordinates' types before they contract with him. The timing of events therefore is, first, the lowest-tier principals contract with their agents; then the second-lowest principals contract with the lowest principals; and so on up the hierarchy.

Ultimately, the top principal will want agent $i$ to produce an output $Q_i$ as a function of the types. These types will not be revealed to him, however, until after he announces the functions $Q_i$. At the time the lower-tier contracts are being written, this output function has not been set. Each lower-tier principal therefore writes a complete contingent contract; payment is made contingent not only on the vector of reported types $t^r=(t^r_1, \ldots, t^r_n)$ and the realized output $q_i$, but also on the functions $Q_i$ that a middle-tier principal will be asked by the top principal to implement. We shall establish, by induction, the multi-level generalization of equation (4). Consider the decision of a $(j+l)$th-tier principal, for $l<j+l<k$ (i.e., any principal other than the top principal). The expected payment by a $(j+l)$th-tier principal to a $j$th-tier principal will be shown to satisfy (with $I(j)$ indexing the agents below this principal and $E_I$ denoting the expectation over the vector of agents' types):

(5) $\Sigma_{i\epsilon I(j)}\ E_t\ [C^{j+1}(Q_i(t), t_i)] = \Sigma_{i\epsilon I(j)}\ E_t\ [C^j(Q_i(t), t_i) - h(t_i)C^j_t(Q_i(t), t_i)].$

This is established by an argument that mimics that given in Section 2, but with added notational complexity. For any feasible output profile desired by the top principal, $Q_i$, the $j$th-tier subprincipal controlling agent $i$ is offered a payment $P^j_i(q_i, Q_i, t^r)$ for $i$'s output. For any $Q_i$ that may arise in equilibrium, the function $P^j_i$ can be assumed to induce incentive compatibility without loss of generality (Myerson, 1982). For any given $Q_i$, the $j$th-tier principal earns a profit $\pi^j_i(t_i)$ from the output of agent $i$ below him, with

(6) $$\pi^j_i(t_i) = \max_{q_i,\ t^r_i}\ E_{-i}\ [P^j_i(q_i, Q_i, t_{-i}, t^r_i) - C^j(q_i, t_i)],$$

where $t_{-i} = (t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n)$. Using the Envelope Theorem, evaluating derivatives at truth-telling (i.e., $t^r_i = t_i$), we get (denoting by $E_{-i}$ expectations taken over $t_{-i}$):

(7) $$\frac{d\pi^j_i}{dt_i} = \frac{\partial\pi^j_i}{\partial t_i} = -E_{-i}\ [C^j_t(Q_i(t), t_i)].$$

(Notice we now equate actual output $q_i$ to the top principal's desired output $Q_i(t)$, because (7) is evaluated at equilibrium.) Thus the sum of the expected payments by the $(j+1)$th-tier principal to this $j$th-tier principal is

(8) $$\Sigma_{i\epsilon I(j)}\ P^j_i(Q_i(t), Q_i, t) = \Sigma_{i\epsilon I(j)}\ E_t\ [C^j(Q_i(t), t_i) + \pi^j_i(t_i)]$$
$$= \Sigma_{i\epsilon I(j)}\ E_t\ [C^j(Q(t), t_i) - h(t_i)C^j_t(Q_i(t), t_i)].$$



(The second step uses (7), integration by parts, and individual rationality, that is, $\pi^j_i(0)\geq 0$.) Hence we have derived the induction formula (5). The $(j+1)$th-tier principal must offer an informational rent of $h(t_i)C^j_t(Q_i(t), t_i)$ to the $j$th-tier principal because of the private information that subprincipal has obtained from each agent $i$ below him.[9]

The lowest-tier principal has no private information of his own; but he inherits his agent's private information. Thus the hierarchy magnifies the informational rents. How quickly do informational rents increase as we move up the hierarchy? By examining equation (5), we see that the $j$th-tier information cost depends on the $j$th derivatives with respect to $t$ of $C^0(q, t)$ and $[1-F(t)]/f(t)$, so little can be said in general. In a tractable special case, however, the information costs rise surprisingly quickly. Let $F$ be the uniform distribution on [0, 1] and $C^0(q, t)$ take the form $(z+1-t)c(q)$, with $z\geq 0$, (thus making the higher derivatives zero). For this example, substitution in (5) shows that the cost effectively borne by the $j$th-tier principal is $[z+2^j(1-t)]c(q)$. Thus each layer added to the hierarchy doubles the information cost borne by the overall principal. This is because each extra layer adds an information cost equal to the previous layer's information cost. With this functional form, $z$ parameterizes the relative importance of production costs and information costs. $z$ is the cost factor common to all types of agents; the closer $z$ is to zero, the more idiosyncratic are the agents' production costs, and the more important are the information costs. In the limiting case of $z=0$, the cost effectively borne by the $j$th-tier principal is $2^j$ times the actual production cost.

## 6. Contractual Incentives within the Hierarchy

Consider the simple hierarchy discussed above, consisting of a single principal and a single agent. As we saw in Section 2, a sufficient condition for the first-order

conditions to characterize the solution is $C^1_{qt} \leq 0$ and $C^1_{qq} \geq 0$. Now add an extra layer of hierarchy, so there is a principal, a subprincipal, and an agent.[10] Let $q^2(t)$ be the top (second-tier) principal's desired output function, taking into account the cumulative information costs that the top principal bears.

Does adding an extra layer to the hierarchy result in an extra output distortion? Given the concavity of the total-revenue function, it does if the top principal's marginal cost is higher in the two-tier hierarchy than in the one-tier hierarchy. The marginal cost borne by the top principal in the two-tier hierarchy is $C^2_q = C^1_q - h(t)C^1_{qt}$, and so $C^2_q \geq C^1_q$ if and only if $C^1_{qt} \leq 0$, which we have already seen is part of the sufficient condition for the two-tier hierarchy to be implementable. Hence output falls as the hierarchy lengthens.

By induction, a $k$-tier hierarchy is workable if and only if $dq^k(t)/dt \geq 0$; and a sufficient condition for this is $C^j_{qq}(q(t), t) \geq 0$ and $C^j_{qt}(q(t), t) \leq 0$ for all $j$, $0 \leq j \leq k$, and for all $t$. Since $C^j_{qq}$ and $C^j_{qt}$ depend on $(j+2)$th derivatives of $C^0$, this is a strong and uninterpretable requirement.

With an optimal linear contract, the marginal payment rate for a $j$th-tier principal in a $k$-tier hierarchy is $C^j_q(q^k(t), t)$, where $q^k(t)$ denotes the output ordered by the top ($k$th) principal. But this marginal payment rate is equal to $C^{j-1}_q(q^k(t), t) - h(t)C^{j-1}_{qt}(q^k(t), t)$. Given, as before, that $C^{j-1}_{qt} \leq 0$, this means that the marginal rate of payment rises as we move up the hierarchy: a supervisor's performance bonus always exceeds his supervisee's. If a supervisor controls $m$ agents, then the bonus the supervisor receives from his superior is more than $m$ times the bonus given to any of the agents.

A reduction in the desired output reduces the marginal rate of payment to an agent (that is, $C^0_q(q^k(t), t)$). Thus the longer the hierarchy, the smaller the agent's marginal payment rate, given the agent's type. Small firms will tend to pay workers piece rates; large firms' workers' payments will look more like wages. In larger firms, workers will tend to receive more of their utility in the form of nonpecuniary benefits



like leisure. Because of the unobservability of the agents' actions, efficient production requires that the agents be paid at a marginal rate of one. The longer the hierarchy above an agent, the larger the effort distortion. Large firms produce less efficiently than small firms.

The measure of efficiency of an agent in a $k$-tier hierarchy is the ratio of lowest-possible cost to actual cost, or $C^0(q, t)/C^k(q, t)$. As the hierarchy lengthens, $C^k(q, t)$ rises and efficiency declines. Thus we see that the degree of efficiency varies not only with external factors but also with the firm's internal organization.

To illustrate these points, let the agents' production cost be $\{(z+1-t)q^2\}/2$, and consider the effects of the hierarchy on one particular agent with type $t$. The effective cost borne by a $j$th-tier principal for this agent's output is $C^j(q, t)$, or (from (5)) $([z+2^j(1-t)]q^2\}/2$. The top, or $k$th-level, principal, has a monopoly on the sale of the output, and demand is linear, with price equal to $a-bq$. The principal maximizes $q(a-bq)-\{[z+2^k(1-t)]q^2\}/2$. Thus the output the principal wants is $a/[2b+z+2^k(1-t)]$. Hence the agent's output falls as the length of the hierarchy above him, $k$, rises. The principal evokes this output with a contract offering a marginal payment rate of $C^j_q(q^k(t), t)$, or $a[z+2^j(1-t)]/[2b+z+2^k(1-t)]$. Compare these with the full-information output, which is $a/[2b+z+(1-t)]$, and the full-information marginal payment rate, which is one. Except in the extreme case in which the agent has maximum productivity ($t=1$), (a) any player's marginal rate of payment is strictly less than one, so the induced effort is less than at the full-information optimum; (b) for a player at a given level in a hierarchy, the marginal rate of payment falls as the total length of the hierarchy increases; and (c) within a given hierarchy, the marginal rate of payment rises as we move up the hierarchy. (It can be shown that $dC^j_q(q^k(t), t)/dt$ is positive with these functional forms provided $b$ is small enough relative to $a$, specifically $2b<az$. Hence linear contracts do indeed work at each level of the hierarchy.)

## 7. Conclusion

We have proposed strategic behavior in heriarchies as a source of organizational diseconomies of scale. When information about the firm's capabilities is dispersed among the individuals in the firm, production is inefficient even though all the firm's members behave rationally.

As a description of a firm, this model has some obvious shortcomings, the correction of which is left for future research. First, for the foregoing analysis to apply, the uncertainties the different workers face must be uncorrelated. If, more realistically, correlations exist then the principal can use relative performance evaluations to mitigate his informational disadvantages. Second, all but the bottom members of the hierarchy are modeled as being relatively passive; they simply control the people immediately below them. A more satisfactory model would have people in the middle of the hierarchy explicitly modeled as being engaged in production, allowing comparisons of the transactions costs of internal and external vertical exchanges. Third, although we have interpreted the hierarchy as being within a firm, it could also be interpreted as a hierarchy of separate subcontractors linked by incentive contracts, on the Japanese model. For our model to distinguish between these two organizational modes, the possibility of contractual incompleteness might have to be added. Fourth, we have examined only the costs of operating a hierarchy. In order to explain why the hierarchy exists, the model would have to be expanded to include some notion of a bounded managerial span of control.



## References

Calvo, Guillermo A., and Wellisz, Stanislav, "Supervision, Loss of Control, and the Optimum Firm Size," *Journal of Political Economy* 86 (1978): 942-53.

Demski, Joel S., and Sappington, David E. M., "Hierarchical Regulatory Control," *RAND Journal of Economics* 18 (Autumn 1987): 369-83.

Farrell, M. J., "The Measurement of Productive Efficiency," *Journal of the Royal Statistical Society* 120, Part III (1957): 253-281.

Galbraith, J. K., *The New Industrial State*, 3rd edn. New York: New American Library, 1979.

Geanakoplos, John, and Milgrom, Paul, "A Theory of Hierarchies Based on Limited Managerial Attention," unpublished, Yale University, October 1985.

Guesnerie, Roger, and Oddou, Claude, "Increasing Returns to Size and their Limits," *Scandanavian Journal of Economics* 90 (1988): 259-73.

Hart, Oliver D., "The Market Mechanism as an Incentive Scheme," *Bell Journal of Economics* 14 (Autumn 1983): 366-82.

Hayek, F. A., "The Use of Knowledge in Society," *American Economic Review* 35 (September 1945): 519-30.

Hicks, John R., "Annual Survey of Economic Theory: The Theory of Monopoly," *Econometrica* 3 (January 1935): 1-20.

Holmström, Bengt, and Tirole, Jean, "The Theory of the Firm," in *Handbook of Industrial Organization*, edited by R. Schmalensee and R. Willig. Amsterdam: North-Holland, 1989.

Itoh, Hideshi, "Collusion, Incentives, and Risk Sharing," unpublished, University of California, San Diego, April 1989.

Jensen, Michael, and Meckling, William, "Theory of the Firm: Managerial Behavior, Agency Costs, and Ownership Structure," *Journal of Financial Economics* 3 (October 1976): 305-60.

Keren, M., and Levhari, D., "The Internal Organization of the Firm and the Shape of Average Costs," *Bell Journal of Economics* 14 (1983): 474-86.

Laffont, Jean-Jacques (1988), "Hidden Gaming in Hierarchies: Facts and Models," *Economic Record* 64 (December 1988): 295-306.

Laffont, Jean-Jacques, and Tirole, Jean, "Using Cost Observation to Regulate Firms," *Journal of Political Economy* 94 (June 1986): 614-41.

Leibenstein, Harvey, "Allocative Efficiency vs. 'X-Efficiency'," *American Economic Review* 56 (June 1966): 392-415.

Machlup, Fritz, "Theories of the Firm: Marginalist, Behavioral, Managerial," *American Economic Review* 57 (1967): 1-33.

Mailath, George, and Postlewaite, Andrew, "Asymmetric Information Bargaining Problems with Many Agents," CARESS Working Paper #88-10, University of Pennsylvania, June 1988.

McAfee, R. Preston, and McMillan, John, "Competition for Agency Contracts," *RAND Journal of Economics* 18 (Summer 1987): 296-307.

Melamud, Nahum, Mookherjee, Dilip, and Reichelstein, Stefan, "Hierarchical Decentralization of Incentive Contracts," unpublished, Stanford University, August 1989.

Milgrom, Paul R., "Employment Contracts, Influence Activities, and Efficient Organization Design," *Journal of Political Economy* 96 (February 1988): 42-60.

Milgrom, Paul R., and Roberts, John, "An Economic Approach to Influence Activities in Organizations," *American Journal of Sociology* 94, Supplement (1988): S154-S179.

Milgrom, Paul R., and Roberts, John, "Bargaining and Influence Costs and the Organization of Economic Activity," in *Rational Perpectives on Political Economy*, edited by J. Alt and K. Shepsle (Cambridge: Cambridge University Press, 1989).

Mirrlees, James A., "The Optimal Structure of Incentives and Authority within an Organization," *Bell Journal of Economics* 7 (Spring 1976): 105-31.

Myerson, Roger B., "Optimal Coordination Mechanisms in Generalized Principal-Agent Problems," *Journal of Mathematical Economics* 10 (1982): 67-81.

Rogerson, William P., "On the Optimality of Menus of Linear Contracts," unpublished, Northwestern University, March 1987.

Rosen, Sherwin, "Authority, Control, and the Distribution of Earnings," *Bell Journal of Economics* 13 (1982): 311-23.

Samuelson, Paul A., *Economics*, 10th edn. New York: McGraw-Hill, 1976.

Scharfstein, David, "Product Market Competition and Managerial Slack," *RAND Journal of Economics* 19 (Spring 1988): 147-55.

Scherer, F. M., *Industrial Market Structure and Economic Performance*. Boston: Houghton Mifflin, 2nd edn., 1980.

Schiff, Michael, and Lewin, Arie Y., "The Impact of People on Budgets," *Accounting Review* 45 (April 1970): 259-68.

Smith, Adam, *An Enquiry into the Nature and Causes of the Wealth of Nations*. Chicago: University of Chicago Press, 1776, 1976.

Spulber, Daniel F., "Incentives and Costs in the Competitive Firm," unpublished, University of Southern California, July 1989.

Tirole, Jean, "Hierarchies and Bureacracies: On the Role of Collusion in Organizations," *Journal of Law, Economics, and Organization* 2 (Fall 1986): 181-214.



Williamson, Oliver E., "Hierarchical Control and Optimum Firm Size," *Journal of Political Economy* 75 (April 1967): 123-38.

Williamson, Oliver E., *The Economic Institutions of Capitalism*. New York: Free Press, 1985.

**Footnotes**

* We thank Bengt Holmström, Hideshi Itoh, Eric Rasmussen, Michael Rothschild, and seminar participants at Berkeley, Columbia, UCLA, UCSD, Yale, and the Decentralization Conference at Caltech for useful comments.

1. In, respectively, *Business Week*, March 27 1989, pp. 84-94; *New York Times* September 24 1989, p. F3; *World Trade* Summer 1989, pp. 34-68; and *Far Eastern Economic Review* August 17 1989, p. 80.

2. Melamud, Mookherjee, and Reichelstein (1989) also model asymmetric information as a source of hierarchy costs; a comparison with their model will be made below.

3. How is the agent's understating of costs consistent with the argument leading to equation (4), in which agents are envisaged as telling the truth? We have seen that, to induce truth-telling, the principal must offer the agent some rents. An equivalent but more realistic process has the principal accepting the agent's reports at face value, knowing they involve some exaggeration, and again letting the agent keep some rents.

4. Hart (1983) develops a model in which firms with separate owners and managers compete with owner-managed firms. Under the assumption that managers are very risk averse, Hart shows that managerial slack is lowered by the existence of competition. Scharfstein (1988), however, shows that with less extreme risk aversion competition may increase managerial slack.

5. Spulber (1989) develops a related model of the internal organization of a perfectly competitive firm, and looks at the interactions between agent effort decisions and purchased-input quantities.

6. This cost of merger could be avoided if the principal could commit in advance of the merger not to extract rents by introducing the inefficiency; the implicit assumption therefore is that such a commitment is not feasible.

7. Our model assumes away the possibility that lower-level players might collude against the top principal; for models of collusion in hierarchies, see Itoh (1989), Laffont (1988), and Tirole (1986).

8. Demski and Sappington (1987) model a three-level hierarchy with adverse selection and moral hazard. Unlike in our model, the top principal designs all of the contracts. The intermediate principal is able to gather improved information about the agent's productivity. Thus there is an extra level of moral hazard: the top principal must motivate the intermediate principal to acquire the information. He does this by introducing extra distortions in the agent's payment schedule.

9. Consider a different sequence of events. Suppose the top principal can prevent the lower principals from offering their contracts until after he has contracted with them, so the top principal contracts first, then the $(k-1)$th-tier principals, and so on down to the lowest principals. Then at the time they accept their contracts the intermediate principals have no private information and earn no informational rents. Only the agents earn informational rents, and it is as if the hierarchy has only one stage. Melamud, Mookherjee, and Reichelstein (1989) model a three-level hierarchy in which the middle agent as well as the bottom agent have private information. Contracts are



written from the top down; unlike in the present model, however, this does not collapse to effectively one level of hierarchy, because of the private information in the middle of the hierarchy.

10. We could make the hierarchy less trivial, as in the previous section, by having more than one person at each of the two lower tiers. But because of the assumed independence of agents' types, this would add only notational complication.

Appendix: Derivations (not to be published)

The price is $p(Q)=a-bQ$, $cost=zq+(1-t)q^2$.

Let $\alpha \equiv a-z$ $z$, $a$ chosen so that price never goes negative. $\beta=b^{-1}$.

*Competition*

The $i^{th}$ firm's profits are

$$\pi_i=E_t q_i(a-bQ)-zq_i-(1-t)q_i^2$$
$$=[\alpha-b(n-1)\mu]q_i-(b+(1-t_i))q_i^2,$$

where $\mu=Eq_i^*(t_i)$. Thus

$$q_i^*(t_i)=\frac{1}{2}\frac{\alpha-b(n-1)\mu}{b+1-t_i}$$

$$\mu=Eq_i^*(t_i)=\frac{1}{2}(\alpha-b(n-1)\mu)E\frac{1}{b+1-t}$$
$$=\frac{1}{2}(\alpha-b(n-1)\mu)\log(1+\beta),$$

where,

$$E\frac{1}{b+1-t}=\int_0^1\frac{dt}{b+1-t}=-\log(b+1-t)\Big|_0^1=\log(b+1)-\log(b)=\log(1+b^{-1})=\log(1+\beta)$$

Thus,

$$\mu[2+b(n-1)\log(1+\beta)]=\alpha\log(1+\beta),\ \text{or},$$

$$\mu=\frac{\alpha\log(1+\beta)}{2+b(n-1)\log(1+\beta)}.$$

$$E\pi_i=E\frac{1}{4}\frac{(\alpha-b(n-1)\mu)^2}{b+1-t_i}=\frac{1}{4}(\alpha-b(n-1)\mu)^2E\frac{1}{b+1-t}$$

$$=\frac{1}{4}\left[\frac{2\mu}{\log(1+\beta)}\right]^2=\frac{\alpha^2\log(1+\beta)}{(2+b(n-1)\log(1+\beta))^2}$$

Industry Profits under competition are

$$\pi^c=nE\pi=\frac{n\alpha^2\beta^2\log(1+\beta)}{[2\beta+(n-1)\log(1+\beta)^2]}=\frac{n\alpha^2\log(1+\beta)}{[2+(n-1)\frac{1}{\beta}\log(1+\beta)]^2}$$

$$\lim_{n\to\infty}\pi^c=0.$$

$$\lim_{\beta\to 0}\pi^c=0.$$

$$\lim_{\beta\to\infty}\pi^c=\infty.$$

$$\lim_{\beta\to\infty}\frac{\pi^c}{\log(1+\beta)}=\frac{n\alpha^2}{4}.$$

$$\lim_{\beta\to 0}\frac{d\pi^c}{d\beta}=\frac{n\alpha^2}{(n+1)^2}.$$

Thus, for small $\beta$,

$$\pi^c\sim\frac{n\alpha^2}{(n+1)^2}\beta.$$

*Monopoly*

$$y\equiv\sum_{i=1}^{n}(1-t_i)^{-1}.\qquad \textit{Note}\ \ y\geq n,\ Ey=\infty.$$

The principal's profits are:

$$\pi^1=(\alpha-b\Sigma q_i)\Sigma q_i-\Sigma 2(1-t_i)q_i^2.$$

$$0=\frac{\partial\pi^1}{\partial q_i}=\alpha-2bQ-4(1-t_i)q_i.$$

$$q_i^*=\frac{\alpha-2bQ}{4(1-t_i)}.$$

$$4Q=(\alpha-2bQ)\Sigma(1-t_i)^{-1}=(\alpha-2bQ)y.$$

$$Q[4+2by]=\alpha y,\ \text{or},\qquad Q=\frac{\alpha y}{4+2by}.$$

$$\alpha-bQ=\frac{\alpha(4+2by)-b\alpha y}{4+2by}=\alpha\frac{4+by}{4+2by}.$$

$$\alpha - 2bQ = \frac{\alpha(4+2by) - 2b\,\alpha y}{4+2by} = \frac{4\alpha}{4+2by}.$$

Thus,

$$q_i^* = \frac{\alpha}{(1-t_i)(4+2by)}, \text{ and },$$

$$\pi^1 = (\alpha - bQ)Q - \Sigma 2(1-t_i)q_i^2$$

$$= \alpha\left[\frac{4+by}{4+2by}\right]\frac{\alpha y}{4+2by} - 2\Sigma\frac{\alpha^2}{(1-t_i)(4+2by)^2}$$

$$= \frac{\alpha^2}{(4+2by)^2}\left[(4+by) - 2y\right]$$

$$= \frac{\alpha^2}{4}\frac{y}{2+by} = \frac{\alpha^2}{4}\frac{\beta y}{2\beta + y}$$

As $n \to \infty$, $y \to \infty$ (since $y \geq n$)

$$\pi^1 = \frac{\alpha^2}{4}\frac{1}{b+2/y} \to \frac{\alpha^2}{4b} \quad as \quad n \to \infty.$$

$$\pi^1 = \frac{\alpha^2}{4}\frac{\beta y}{2\beta + y} \to 0 \quad as \quad \beta \to 0.$$

$$\lim_{\beta \to 0}\frac{d}{d\beta}\pi^1 = \lim_{\beta \to 0}\frac{\alpha^2}{4}\frac{y(2\beta+y) - 2\beta y}{(2\beta+y)^2}$$

$$= \frac{\alpha^2}{4}\lim_{\beta \to 0}\frac{y^2}{(2\beta+y)^2} = \frac{\alpha^2}{4}$$

Thus, for small $\beta$,

$$\pi^1 > \pi^c \text{ IFF } \frac{n}{(n+1)^2} < \frac{1}{4}, \text{ which is true for } n \geq 2.$$

TOTAL MONOPOLY PROFITS

If, in order to buy the competitive firms, the metaprincipal must only pay the difference between what agents expect under competition and what they expect under the metaprincipal, the total profits of the monopoly are the relevant comparison to $\pi^c$ (i.e. is there anything left over for the



metaprincipal?). These profits are

$$\pi^m = (\alpha - bQ)Q - \Sigma(1-t_i)q_i^2$$

$$= \alpha\frac{4+by}{4+2by}\frac{\alpha y}{4+2by} - \Sigma(1-t_i)\frac{\alpha^2}{(1-t_i)^2(4+2by)^2}$$

$$= \frac{\alpha^2}{(4+2by)^2}\left[y(4+by) - y\right]$$

$$= \frac{\alpha^2(3+by)y}{(4+2by)^2} = \frac{\alpha^2}{4}\frac{3y+by^2}{4+4by+b^2y^2}$$

$$= \frac{\alpha^2}{4}\frac{3y^{-1}+b}{4y^{-2}+4by^{-1}+b^2}$$

As noted earlier,

$$\frac{n\alpha^2}{4} = \lim_{\beta \to \infty}\frac{\pi^c}{\log(1+\beta)} = \lim_{\beta \to 0}\frac{\pi^c}{\log(1+b^{-1})}$$

$$\lim_{b \to 0} E\pi^m = E\lim_{b \to 0}\pi^m = \frac{3}{16}\alpha^2 Ey = \infty$$

Although y is an improper random variable (no mean), $y^{-1}$ is not. Since $y \geq n$, $y^{-1} \in [0, 1/n]$.

Moreover $Ey^{-1} > 0$, since there is a positive probability that $y \leq 2n$.

$$\lim_{b \to 0}\frac{E\pi^m}{\log(1+b^{-1})} = E\lim_{b \to 0}\frac{\pi^m}{\log(1+b^{-1})} = E\lim_{b \to 0}\frac{\frac{\partial}{\partial b}\pi^m}{\frac{\partial}{\partial b}\log(1+b^{-1})}$$

$$= \frac{\alpha^2}{4}E\lim_{b \to 0}\frac{4y^{-1}+b}{(2y^{-1}+b)^3} \Big/ \frac{1}{b(b+1)}$$

$$= \frac{\alpha^2}{4}E\lim_{b \to 0}\frac{(4y^{-1}+b)b(b+1)}{(2y^{-1}+b)^3} = 0$$

Thus $\lim_{b \to 0}\frac{E\pi^m}{\pi^c} = 0$, and thus, for small b:

$$E\pi^1 \leq E\pi^m < \pi^c$$

This gives the following summary:

(i) For $n$ large, $\pi^m \geq \pi^1 > \pi^c$

(ii) For $b$ large, $\pi^m \geq \pi^1 > \pi^c$

(iii) For $b$ small, $\pi^1 \leq \pi^m < \pi^c$