Organizational Fixed Costs and Organizational Structure*

Benjamin E. Hermalin

October 31, 1996

Abstract

This paper shows how fixed costs can play an important role in determining the structure of organizations. Applications considered include layoffs, downsizing, heterogeneity among firms, and the structure of hierarchies. Agency problems as a source of fixed costs are also considered.
Keywords: *fixed costs, organizational structure.*

1 Introduction

The importance of fixed costs for explaining industry structure is well known. For instance, it won't work to assume that firms' cost functions are q^2 in a competitive model—if literally true, then free entry would lead to an arbitrarily large number of firms each producing an arbitrarily small amount of output. That is, there would be no equilibrium structure. A fixed-cost assumption prevents this preposterous outcome: For example, assume that the firms' cost functions are $q^2 + F$ (for $q > 0$). Average cost is, then, the familiar U-shape shown in microeconomic texts. Moreover, a determinant industry structure emerges: If market demand is, say, $A - Bp$, then there are

$$\text{int}\left[\frac{A - B\sqrt{F}}{\sqrt{F}}\right]$$

firms (where $\text{int}[\cdot]$ means the "integer part of").

This relationship between fixed costs and industry structure has an analogy in the study of *organizational* structure. The people (alternatively, departments, plants, divisions, etc.) within an organization are analogous to firms in an industry. The structure of the organization—e.g., the number of people within it and their hierarchical organization—are, then, a consequence of fixed costs within the organization.

To make this analogy more concrete, suppose a firm can hire any number of workers. Each worker has a cost of effort $c(a) = a^2$, where a is units of effort. A worker's overall utility is given by $w - c(a)$, where w is his wage. Assume that effort is observable, so contracts can be contingent on effort. Assume that the firm possesses the bargaining power when hiring workers, so each worker's participation (individual rationality) constraint,

$$w - c(a) \geq \bar{w},$$

*Helpful comments from Bob Anderson, David I. Levine, Tom Marschak, Ilya Segal, and seminar participants at U.C. Berkeley are gratefully acknowledged. Financial support from the U.C. Berkeley Committee on Research, the Walter A. Haas School of Business, and the Harold Furst Chair in Management Philosophy and Values is also gratefully acknowledged.

is binding ($\bar{w}$ is a worker's reservation wage). The firm's cost of getting a units of effort from a worker is, therefore,

$$c(a) + \bar{w} = a^2 + \bar{w}.$$

From this last expression, it is clear that $\bar{w}$ is the fixed cost of effort. Suppose the firm needs an *aggregate* amount of effort A.[1] Since $c(a)$ is convex, it follows that for a *given* number of workers, N, the firm minimizes its costs by dividing the work equally among the workers; that is, having each supply A/N units of effort. The firm's profit, as a function of N, is

$$\beta(A) - N \cdot \left(\left(\frac{A}{N} \right)^2 + \bar{w} \right), \tag{1}$$

where $\beta(\cdot)$ is the firm's benefit from effort function. Suppose, at first, that there were no fixed costs of effort (i.e., $\bar{w} = 0$). Then maximizing (1) with respect to N would be equivalent to minimizing

$$\frac{A^2}{N},$$

which has no determinant solution (i.e., $N \to +\infty$). If, however, there are fixed costs of effort (i.e., $\bar{w} > 0$), then maximizing (1) with respect to N is equivalent to minimizing

$$\bar{w}N + \frac{A^2}{N},$$

which does have a determinant solution:

$$N^* = \frac{A}{\sqrt{\bar{w}}}.$$

Moreover, the importance of a fixed cost to the size of this firm is clear from this last expression: The greater the fixed cost, the fewer workers employed.

Despite the straightforward nature of this analogy, it's a powerful tool for both understanding and developing economic models of organization. The next sections demonstrate this claim. In Section 2, I show how this idea can be employed to explain phenomena like large-scale layoffs and downsizing. It can also explain heterogeneity in size across otherwise identical firms in the same industry. In the following section, I use this idea to build a model that explains why hierarchies tend to be pyramid shaped. In both these sections, the fixed costs are simply assumed. Organizational problems, such as agency problems, can, however, generate fixed costs endogenously. Section 4 develops an agency model that illustrates this point. Section 5 concludes.

2 Determination of Firm Size

As in the previous section, each worker's utility function is $w - c(a)$, where w is wage and $a \in [0, \infty)$ is a measure of the worker's action (e.g., effort). The function $c(\cdot)$ is assumed to be increasing,

[1] Alternatively, one could assume that the individual efforts were subadditive—the benefit of spreading effort among workers is greater than assumed here. Or, one could assume that the individual efforts were superadditive—the benefit of spreading effort among workers is less than assumed here. Provided effort is not too superadditive, the convexity of $c(\cdot)$ assures that effort is evenly divided among workers. Subadditivity lessens the importance of fixed costs, while superadditivity heightens the importance of fixed costs relative to the analysis here.

convex, and, for convenience, at least twice differentiable. Increasing reflects the idea that the worker finds the action distasteful. Convexity is a standard returns-to-scale assumption in organization theory, it indicates that the marginal distaste for the action is increasing in the action. To avoid too many sources of fixed costs, assume that $c(0) = 0$. As before, each worker's participation constraint is

$$w - c(a) \geq \bar{w},$$

where $\bar{w}$ is the reservation wage.[2] Assume that contracts can be contingent on effort (Section 4 relaxes this assumption). Assume, too, that the firm has all the bargaining power. Then the firm can induce a given level of action, $\hat{a}$, from the worker using the forcing contract:

$$w(a) = \begin{cases} \bar{w} + c(\hat{a}) & \text{if } a = \hat{a} \\ -\infty \text{ if } a \neq \hat{a} \end{cases}.$$

It follows that the firm's equilibrium cost of getting a worker to supply action a is, thus, $\bar{w} + c(a)$.

Suppose that the firm wishes to induce an *aggregate* level of action, A, from its N workers. Let $\beta(A)$ be the firm's benefit from aggregate action A. For convenience, assume that $\beta(\cdot)$ is at least twice differentiable. Since $c(\cdot)$ is convex, it follows from Jensen's inequality that the firm minimizes cost by dividing A evenly among the N workers. Conditional on employing N workers, its cost of inducing A is

$$C(A; N) = N\bar{w} + Nc\left(\frac{A}{N}\right).$$

Lemma 1 *For $M > N$, $C(\cdot; M)$ crosses $C(\cdot; N)$ once from above.*[3]

Figure 1 illustrates this result.

The number of workers, N, is endogenous, so the *un*conditional cost of inducing A is

$$C^*(A) = \min_N C(A; N).$$

Provided $\bar{w} > 0$, this minimization problem is well defined and has a finite solution for N. Graphically, $C^*(\cdot)$ is the lower envelope of $\{C(\cdot; N)\}_{N=1}^{\infty}$. This is shown in Figure 2. As the figure makes clear, $C^*(\cdot)$ is *not* convex in the neighborhoods where the $C(\cdot; N)$ curves cross. This non-convexity will be important in the analysis that follows. In particular, it follows that:

Proposition 1 *Define A_N to be where $C(\cdot; N)$ and $C(\cdot; N+1)$ cross. Then there is a neighborhood, $\mathcal{N}$, around A_N such that* no *$A \in \mathcal{N}$ maximizes the firm's profit.*

Proposition 1 means that there are "gaps"—the neighborhoods $\mathcal{N}$—in the set of A's that could maximize profits. Figure 3 illustrates this.

How large might these gaps be? The answer depends on the concavity of the benefit function, $\beta(\cdot)$. The more concave it is, the smaller the gaps. Conversely, if $\beta(\cdot)$ is affine (linear), then the gaps can be quite large. To study the size of the gaps for an affine benefit function, limit attention to

[2] The reservation wage could represent, for example, the worker's cost, including value of time, from commuting to and from work. Equivalently, $\bar{w}$ could be a fixed cost borne by the firm—for example, the cost of providing the worker fringe benefits. If the organizational unit is greater than a worker (e.g., a shift), then $\bar{w}$ could be the fixed cost of organizing and equipping that unit.

[3] Proofs can be found in the appendix.

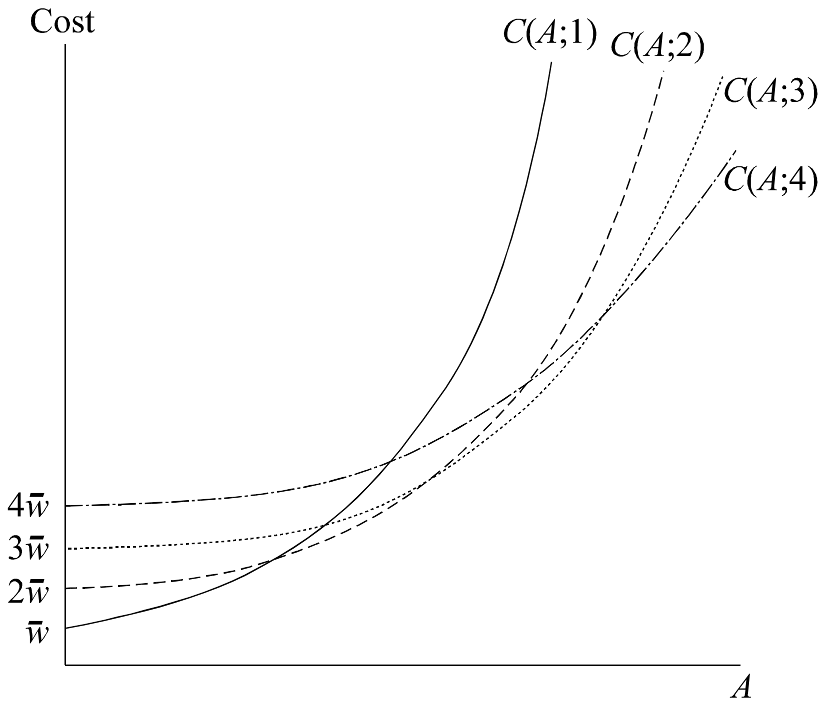

Figure 1: $C(A;M)$ crosses $C(A;N)$ once from above $(M > N)$.

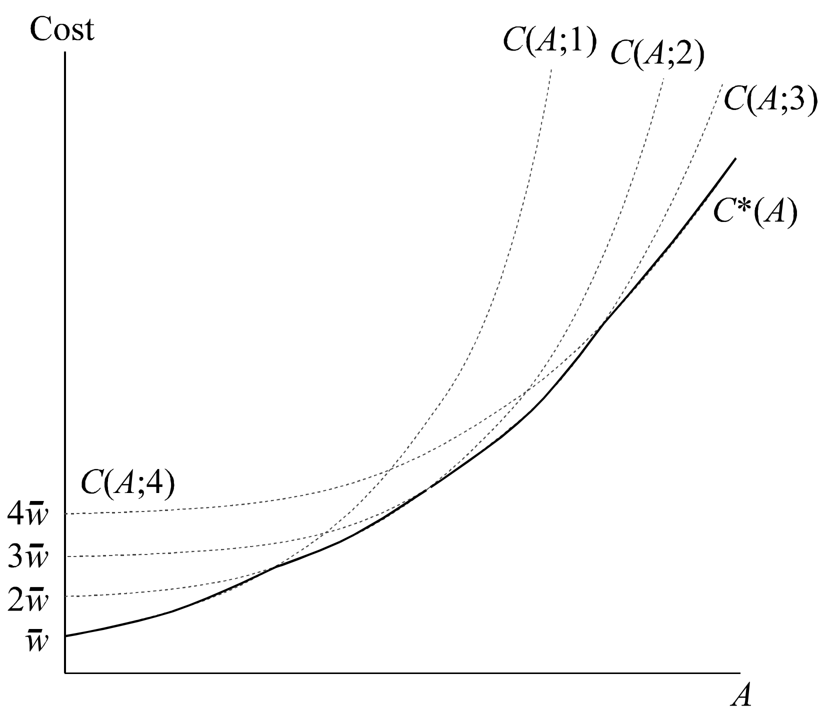

Figure 2: $C^*(A)$ is the lower envelope of the $C(A;N)$.

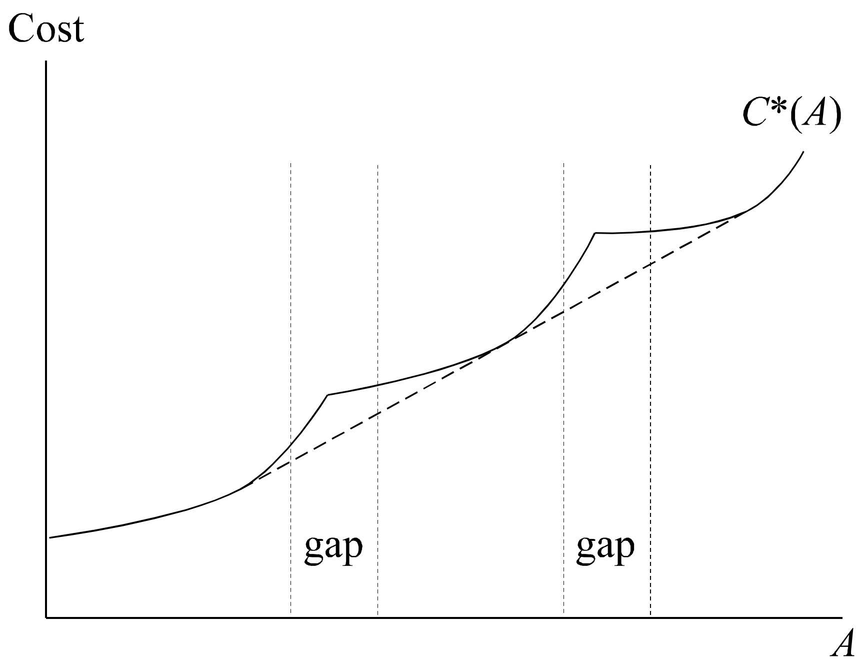

Figure 3: A's in the "gaps" are never optimal.

$N \leq \bar{N}$, where $\bar{N}$ is some maximum level of employment. Let $\bar{A}$ be some corresponding maximum level of aggregate effort. Define $C^{**}(\cdot)$ to be the convex lower bound of $C^{*}(\cdot)$ (sometimes also called the largest convex minorant of $C^{*}(\cdot)$).[4] In Figure 3, $C^{**}(\cdot)$ either coincides with $C^{*}(\cdot)$ or it is the dashed segment below $C^{*}(\cdot)$. The following lemma is useful for defining the gaps for an affine benefit function.

Lemma 2 *Assume the benefit function, $\beta(\cdot)$, is affine. Define the *-program to be*

$$\max_{A \in [0,\bar{A}]} \beta(A) - C^{*}(A).$$

*Define the **-program to be*

$$\max_{A \in [0,\bar{A}]} \beta(A) - C^{**}(A).$$

*Then, if $\hat{A}$ is a solution to the **-program, it is a solution to the *-program if and only if $C^{*}\left(\hat{A}\right) = C^{**}\left(\hat{A}\right)$. Moreover, if $\hat{A}$ is a solution to the *-program, then $C^{*}\left(\hat{A}\right) = C^{**}\left(\hat{A}\right)$.*

In other words, when $\beta(\cdot)$ is affine, the gaps are the set $\{A | C^{*}(A) > C^{**}(A)\}$. How big might this set be? The following proposition answers:

[4] Formally, $C^{**}(A) = \inf\{\mu | (A,\mu) \in \mathrm{co}(\mathrm{epi}(C^{*}))\}$, where $\mathrm{epi}(C^{*})$ is the epigraph of C^{*} (i.e.,

$$\mathrm{epi}(C^{*}) = \left\{(A,\mu) \,|\, A \in \left[0,\bar{A}\right] \text{ and } \mu \geq C^{*}(A)\right\})$$

and $\mathrm{co}(\mathrm{epi}(C^{*}))$ is the convex hull of $\mathrm{epi}(C^{*})$. See van Tiel (1984) for further details.

Proposition 2 *Suppose $\beta(\cdot)$ is affine. Define a^* to be the solution to*

$$ac'(a) = \bar{w} + c(a) \tag{2}$$

and define $\hat{A}_N = Na^*$. Then the gaps are all A *in the interval* $\left(0, \hat{A}_{\bar{N}}\right]$ minus *the finite set* $\left\{\hat{A}_1, \ldots, \hat{A}_{\bar{N}}\right\}$.

That is, if the benefit function, $\beta(\cdot)$, is affine, then *almost every* A between $\left(0, \hat{A}_{\bar{N}}\right)$ is in a gap.

These gaps are important for two reasons. First if $\mathcal{A}$ is the set of A's that could ever be optimal to induce, then the existence of gaps means $\mathcal{A}$ is *non-convex.* This has important implications in strategic settings, because it means that even if firms are otherwise identical, the only pure-strategy equilibria can still be asymmetric. As I argue in Section 2.2, this can help to explain *heterogeneity* in organizational size (also see Hermalin, 1994, for a related analysis).

The gaps are also important for understanding organizational comparative statics. To see why, let θ be a parameter that changes the benefit function. As θ changes, the optimal A, A^*, to induce will also change. The change in A^* with θ will not, however, be everywhere continuous: Around gaps, the change in A^* will be discrete—A^* will have to "jump over" the gaps. This has two implications. First, small changes in a firm's environment (e.g., θ) can have dramatic effects on its behavior (e.g., a big jump in A^*). Second, since gaps include the cutoff values, A_N, a small change in the environment can lead to a discrete change in the size of the firm (e.g., from N to $N+1$). The larger are the gaps, the more frequently they are encountered. This means that a sequence of small changes in the environment can lead to a sequence of large changes in the firm's size and its activity.

In the case of an affine benefit function, these changes will be particularly dramatic. Suppose

$$\beta(A) = \begin{cases} \theta A + k, \text{ if } A > 0 \\ 0, \text{ if } A = 0 \end{cases},$$

where $k \geq 0$. Define

$$\underline{N}(k) = \begin{cases} 1, \text{ if } k > \bar{w} \\ 0, \text{ if } k \leq \bar{w} \end{cases}.$$

Then if $\beta(\cdot)$ is the benefit function, the firm will almost surely be of size $\underline{N}(k)$ or size $\bar{N}$:

Proposition 3 *Suppose the benefit function is $\beta(A) = \theta A + k$ and $N \leq \bar{N}$. Then for all values of the marginal benefit parameter, θ, the optimal size of the firm is either $\underline{N}(k)$ or $\bar{N}$, except if $\theta = c'(a^*)$, in which case all firm sizes are optimal (where a^* is defined in Proposition 2). Firm size will be $\bar{N}$ if $\theta > c'(a^*)$ and it will be $\underline{N}(k)$ if $\theta < c'(a^*)$.*

Under the assumptions of Proposition 3, as θ crosses the value $c'(a^*)$, firm size changes dramatically: either jumping from $\underline{N}(k)$ to $\bar{N}$ or from $\bar{N}$ to $\underline{N}(k)$.

2.1 Application: Downsizing and Layoffs

An intriguing feature of organizations is how rapidly they can shrink and, to a lesser extent, grow. What is puzzling about these changes, such as large-scale layoffs, is that they often come in response to rather slow-moving or predictable changes in the economic environment. In terms of the notation from the last section, N changes quickly while θ changes slowly. Organizational fixed costs and the resulting gaps in the optimal actions set, $\mathcal{A}$, offer one answer to this puzzle.

For example, consider a competitive firm that can sell all its output at a price of p per unit. Let the firm's output be A. Assume that $N \leq \bar{N}$. Then in terms of the earlier notation, $\theta = p$ and $\beta(A) = pA$. Sudden layoffs (or expansions) are, then, an immediate corollary of Proposition 3:

Corollary 1 *Consider a competitive firm that can sell its output at a price of p per unit. Let its output be A. Then employment, N, is constant except when price falls below $c'(a^*)$, at which point it falls to zero.*

So price can be steadily falling, but it is only when it crosses $c'(a^*)$ that employment falls. Moreover, when it falls, it falls dramatically from $\bar{N}$ to 0. Conversely, price can be steadily rising, but it is only when it crosses $c'(a^*)$ that employment expands.[5]

Taken *literally*, Corollary 1 is rather extreme—few, if any, firms go from $\bar{N}$ employees to zero. A more sensible interpretation might be that $\bar{N}$ is the number of workers over which the firm has discretion in some intermediate run (e.g., at a given time a factory needs a minimum of m_0 workers, but can employ a maximum of $m_0 + \bar{N}$ workers). Relatedly, the result makes more sense if applied to parts of the firm (e.g., shifts or departments).[6] That is, it could explain eliminating a shift. Or $\bar{N}$ refers to the number of workers at a given level of management and the Corollary describes the decision to eliminate a layer of management. Finally, N can be interpreted as itself counting units that are larger than a single person:

Remark 1 *If one reinterprets N as counting shifts, departments, or plants, or other units of production, then Corollary 1 offers an explanation for* sudden *and* dramatic *movements in their employment.*

An affine benefit function is, admittedly, an extreme assumption in many contexts. Consider, therefore, a monopolist who faces inverse demand $I - Sq$. Continue to assume that output is A. Now $\beta(A) = IA - SA^2$. In this situation, there are still gaps—the size of which depend on S—but because they are smaller, the comparative statics will be less dramatic. However, if S is small, then the gaps are big, and employment will still move "quickly" in response to small changes in I. For instance, if $\bar{w} = 1$, $c(a) = a^2$, and I moves in steps of one from 4 to 50, then employment also moves in steps of one when $S = \frac{1}{2}$. But if $S = \frac{1}{4}$, then employment moves in steps of two (e.g., $N = 16$ when $I = 10$, but $N = 18$ when $I = 11$); if $S = \frac{1}{8}$, then employment moves in steps of four; and if $S = \frac{1}{10}$, then employment moves in steps of five.[7] More generally, if $\beta(\cdot)$ is not too concave (e.g., S is small), then fairly sudden and sizeable movements in employment can be expected as the consequence of small changes in the economic environment.

This "suddenness" is complemented by adjustment costs or lags. Although adjustment costs or lags can, in themselves, explain sudden large movements in employment, their effect is reinforced by the existence of gaps. Dynamically adjusting employment can be compared to choosing where a

[5] Of course, this abstracts away from any uncertainty in price movement or the possibility that there might be time to build (e.g., a firm might *begin* its expansion when $p < c'(a^*)$ in anticipation of continued price increases if it must allot time to build).

[6] For example, New York Times (1996), page 9, gives an example of a department in Chemical Bank that was downsized from 15 to 1. [Since the department did not shut, presumably the benefit function had a positive intercept (i.e., $k > \bar{w}$), see Proposition 3.]

[7] The *Mathematica* program used to derive these results is available from the author upon request.

train should stop on its journey, with each employment level being a potential station. If gaps are large, then there is little benefit to stopping at each intermediate station—since it remains optimal for such a short period of time (small set of parameter values). If its benefit is small enough, the station—employment level—will be bypassed.

Moreover, the role of fixed costs is essential to explaining why adjustments change employment. After all, what the firm wants to adjust is A (collective effort, action, or hours). Absent fixed costs, there is no relationship between A and N. For instance, suppose $\bar{w} = 0$ and, to give the model closure, suppose that no more than $\bar{N}$ workers can be employed. Then the firm is always of size $\bar{N}$—all that changes are A and the amount asked of each worker, $A/\bar{N}$.[8]

In many recent cases of downsizing (see, e.g., New York Times, 1996), the firm sheds workers, but not output or profits. The model developed here can also help in thinking about this kind of downsizing. Typically, in this kind of downsizing, layoffs are associated with switching to an alternative method of production (e.g., automation, outsourcing, over-seas production, etc.). As with other layoffs, one question is why the switch in methods is so sudden and so complete. Why, for instance, don't firms use a slowly adjusting weighted average of the old and new methods?

To model this, assume that the firm can sell all the output it wishes at a price of p per unit, but only up to $\bar{N}a^*$ units. Assume, too, that there is an alternative method of production, in which the firm's cost would be γA. Assume, however, that this method is initially more costly:

$$\gamma > c'(a^*) . \tag{3}$$

The firm's choice of effort under the original technology, A, maximizes

$$pA + p\left(\bar{N}a^* - A\right) - \gamma\left(\bar{N}a^* - A\right) - C^*(A) .$$

This can be rewritten as

$$\gamma A - C^*(A) + \bar{N}a^*(p - \gamma) .$$

The benefit function is affine in A. Moreover, from (3) and Proposition 2, it follows that the optimal $A = \bar{N}a^*$—there is no use of the alternative method—and $N = \bar{N}$.

Now suppose that this alternative method becomes less costly (e.g., the cost of automation falls). There will be no use of this alternative method and employment will stay constant at $\bar{N}$ until the point that $\gamma = c'(a^*)$, at which point employment will suddenly drop to 0 and the use of the alternative method will jump up sharply (from 0 to $\bar{N}a^*$). Production remains constant and profits actually increase. Overall, these results are consistent with the anecdotal evidence reported in New York Times (1996).

Admittedly, it could be a strong assumption to assume that the alternative method of production has linear costs. Suppose, instead, that its cost were $\gamma(A)$, where $\gamma(\cdot)$ is an increasing and convex function. Then $\beta(A) = p\bar{N}a^* - \gamma\left(\bar{N}a^* - A\right)$, which is concave. As the above discussion of the monopolist made clear, the previous analysis will, however, be a reasonable approximation if $\gamma(\cdot)$ is not too convex.

[8] This also explains why the model is *not* being driven by the fact that workers come in discrete units. Fractional workers are, in a sense, possible: If, say, $\hat{a}$ is full-time work, then half a worker corresponds to asking $\frac{1}{2}\hat{a}$ from a worker. What discourages the firm from doing this is that this "fractional" worker still triggers a full amount of the fixed cost, $\bar{w}$.

2.2 Application: Heterogeneity in Organizational Size

A feature common to almost every industry is how its firms vary in size. For instance, General Motors is larger than Ford, which is larger than Chrysler. In this section, I consider how fixed-cost-induced gaps can help to explain this heterogeneity.[9]

Consider a Cournot duopoly (although the reasoning would extend to more firms and alternative forms of competition). Let each firm's benefit function be $q_i\left(I - S\left(q_i + q_j\right)\right)$, where q_i is its own output and q_j is its rival's output. Suppose that $q = A$.

In this model, were there no gaps (i.e., were $C^*(\cdot)$ an everywhere convex function), then no asymmetric pure-strategy equilibrium could exist:

Proposition 4 *Consider a Cournot duopoly with identical firms in which inverse market demand is $I - SQ$, where Q is market output. If $C^*(\cdot)$ were everywhere convex, then there would be a unique pure-strategy equilibrium and it would be symmetric.*

On the other hand, *with* gaps it can be that all pure-strategy equilibria are asymmetric:

Example 1 *Consider a Cournot duopoly with identical firms in which inverse market demand is $5 - \frac{Q}{\sqrt{12}}$, where Q is market output. Assume $\bar{w} = 1$ and $c(a) = a^2$.* Conditional *on employing N workers, firm i's reaction function is*

$$A_i = \frac{N\left(5 - \frac{A_j}{\sqrt{12}}\right)}{2\left(\frac{N}{\sqrt{12}} + 1\right)}. \tag{4}$$

It is readily shown that $A_N = \sqrt{N(N+1)}$ (the cutoff between N and $N+1$ workers). A symmetric equilibrium exists only if (4) has a solution with $A_i = A_j$ and this solution lies between A_{N-1} and A_N. If (4) has a symmetric solution, then

$$A(N) = 10\frac{N}{N\sqrt{3} + 4}.$$

It is readily shown that the $A(N) \in (A_{N-1}, A_N)$ only if $N = 3$ or 4. $A(3) = 3.2622$. However, the best response to $A(3)$ is not to hire 3 workers and induce total effort $A(3)$, but to hire 4 workers and induce total effort 3.7669. $A(4) = 3.6603$. However, the best response to $A(4)$ is not 4 workers and $A(4)$ in effort, but to hire 3 workers and induce total effort 3.1699. Moreover, under these assumptions, it can be shown that $A_i = 3.1374$ and $N_i = 3$ for firm i and $A_j = 3.8004$ and $N_j = 4$ for firm j is a pure-strategy equilibrium.[10]

In Example 1, the difference in employment between the firms is one. Obviously, if applied to large corporations, this would not be a significant source of heterogeneity. On the other hand, if the units are larger than one worker—e.g., a department, plant, or division—then the results are

[9] Admittedly, there are other possible explanations: firms can have access to different production technologies (e.g., because of patents) or they can face different factor prices (e.g., a consequence of producing in different locations). But it is doubtful that these are universal explanations—they do not, for example, seem to fit the domestic automobile industry particularly well. Aron (1988) and Lucas (1978) offer another explanation based on hetorgeneity in managerial ability. These models, however, seem better suited to explain *inter*-industry heterogeneity in size rather than *intra*-industry heterogeneity.

[10] The programs for verifying this are available from the author upon request.

more meaningful. Alternatively, continue to think about the units being workers, but workers at a sufficiently high level of management (e.g., vice presidents), so that a difference of one worker would correspond to one less department, plant, or division.

Could the difference in employment be bigger than one? The answer depends on the size of the gaps, which, in turn, depends on the benefit function. If firm i's benefit function is affine in its own total effort, then the answer is yes. For example, let A_i be effort on advertising and suppose that firm i's revenue is

$$R_0 + \frac{A_i}{\phi(A_j)} \text{ if } A_i > 0$$

and zero otherwise, where $R_0 > \bar{w}$ is revenue from customers unaffected by advertising and $\phi(\cdot)$ is an increasing function. Again assume $N \leq \bar{N}$. Finally suppose

$$\begin{aligned} \frac{1}{\phi(a^*)} &> c'(a^*), \\ \frac{1}{\phi(\bar{N}a^*)} &< c'(a^*), \text{ and} \\ \frac{1}{\phi(Na^*)} &\neq c'(a^*), \text{ for any } N. \end{aligned}$$

Then a pure-strategy equilibrium is $A_i < a^*$ and $N_i = 1$, while $A_j \geq \bar{N}a^*$ and $N_j = \bar{N}$. In fact,

Proposition 5 *For the advertising duopoly game, the only two pure-strategy equilibria have the form $A_i < a^*$ and $N_i = 1$, while $A_j \geq \bar{N}a^*$ and $N_j = \bar{N}$.*

3 Pyramid-Shaped Hierarchies

In this section, I consider how organizational fixed costs can help to explain why hierarchies are pyramid shaped—higher tiers are smaller than lower tiers. There are, admittedly, many facets to hierarchies and many models of hierarchies,[11] so this is not an attempt at a comprehensive or definitive model of hierarchy (if such a creature exists). Rather, the intention is to illustrate further the importance of organizational fixed costs in understanding organizations, while contributing to an understanding of one facet of hierarchies, namely their shape.

A worker's function is to solve problems (e.g., respond to customer complaints, improve production, design corporate strategy, etc.). Let t index problems and let α_t denote effort or attention devoted to problem t. A worker's total effort is a, the sum over his α_t's. As before, his utility is $w - c(a)$. Let $\pi_t(\alpha)$ be the probability that a worker succeeds in solving problem t conditional on devoting α attention to it. Assume that $\pi_t(0) = 0$ and $\pi_t(\cdot)$ is non-decreasing for all t. To make the model more interesting, I focus on problems that are not readily solvable: assume, therefore, that $\pi_t(\cdot)$ has a least upper bound less than one. Let b_t be the benefit of solving problem t.

Suppose that if a worker fails to solve a problem, then the problem is passed up to his superior.[12] His superior's probability of solving it is also $\pi_t(\alpha)$, where α is the attention the superior devotes

[11] A *partial* list of models is Williamson (1967), Calvo and Wellisz (1978), Sah and Stiglitz (1986), Geanakoplos and Milgrom (1991), Bolton and Dewatripont (1994), and McAfee and McMillan (1995).

[12] The direction of transfer is taken as exogenous in this model. In particular, lateral transfers are simply assumed to be impractical. The exogenous direction of transfer is in keeping with papers in the literature (see, e.g., Sah and Stiglitz, 1986).

to it. Likewise, if the superior fails to solve it, she passes it up to her superior, and so forth.[13]

If there were no fixed costs, then the convexity of $c(\cdot)$ entails that each worker would work exclusively on one problem and, for the same reason, each supervisor would have just one person under her in the hierarchy. That is, the hierarchy would be a rectangle rather than a pyramid.

Proposition 6 *Assume no fixed costs. Then, if a problem, t, satisfies*

$$b_t \pi_t(\alpha) - c(\alpha) > 0 \text{ for some } \alpha > 0, \tag{5}$$

it will have a worker and a potentially infinite sequence of supervisors devoted exclusively to it. If condition (5) fails, then no worker will be assigned to the problem.

Note, too, that absent fixed costs the size of the firm is indeterminate (although it does have an expected value).

Now suppose that there are fixed costs. In particular, assume that prior to tackling any problems, the firm has to hire the employees that will solve them. Assume each worker and supervisor has a reservation wage of $\bar{w}$. Since each employee is guaranteed at least $\bar{w}$, a firm with a finite problems will have only a finite number of employees, in contrast to the situation in Proposition 6

Turning to the shape of the hierarchy, assume that the hierarchy consists of layers indexed by m. The bottom layer (workers) is layer one. Assume that an unsolved problem is passed up layer to layer and cannot skip a layer (e.g., it can't go from $m-1$ to $m+1$). Assume, too, that they cannot be dropped—a continually unsolved problem works its way all the way to the top.

Intuitively, if the hierarchy were rectangular, employees at higher layers would have less to do, in expectation, than employees at lower layers because some problems would be solved before they reach the higher layers. Consequently, the marginal cost of increasing the workload of employees at higher layers is relatively low. Moreover, because it would economize on fixed costs, it would pay to increase their workload by "thinning the ranks"; that is, having more than one subordinate refer unsolved problems to an employee in a higher layer. Hence, the structure of the hierarchy becomes pyramidal.

To formalize this intuition in a tractable model assume that

$$\pi_t(\alpha) = \begin{cases} 0, \text{ if } \alpha < 1 \\ \hat{\pi}, \text{ if } \alpha \geq 1 \end{cases}$$

for all t, where $0 < \hat{\pi} < 1$. That is, an employee supplies one unit of attention to each problem assigned to him. Consequently, if he works on T problems assigned to him, then his cost of effort is $c(T)$. Another implication is that the expected benefit supplied by a given layer is independent of its structure (remember that problems cannot be dropped). Consequently, attention can be focused on the cost side of the ledger.

Define $\hat{q}_m = (1-\hat{\pi})^{m-1}$; that is, $\hat{q}_m$ is the probability that a given problem remains unsolved until reaching the mth layer. Let the maximum number of problems that could reach an employee

[13] This corresponds to Sah and Stiglitz (1986)'s "serial processing." Their model and this model differ, however, on a number of grounds. Most importantly, there is no effort in their model and, hence, no issue of compensation and fixed costs.

be T, then the expected cost incurred by employing him is[14]

$$\mathcal{C}_m(T) = \bar{w} + \sum_{t=0}^{T} \binom{T}{t} (1 - \hat{q}_m)^{T-t} \hat{q}_m^t c(t).$$

Lemma 3 *The cost of allowing a maximum of T problems to reach an employee at layer m is decreasing with m. That is, $\mathcal{C}_m(T) < \mathcal{C}_{m-1}(T)$.*

It can now be established that the hierarchy will be pyramidal:

Proposition 7 *The cost advantage of having fewer employees in a layer increases with the height of the layer (i.e., with m).*

This section has shown how organizational fixed costs can explain organizational features such as pyramidal hierarchies. Admittedly, the model developed here is very simple, which limits the contribution it directly makes to a better understanding of hierarchies. On the other hand, the general elements—superiors "complete" the work of subordinates and fixed costs—provide a strong intuitive explanation for a pyramid structure.

4 Agency as a Source of Fixed Costs

So far the source of fixed costs has been exogenous to the model. In this section, I show how agency problems can serve to introduce non-convexities into the cost functions that are similar in impact to the exogenous fixed costs considered so far. That is, organizational problems, such as agency, can be their own source of "fixed costs."

Assume, now, that an agent's utility function is $u(w) - a$, if he is paid w and takes action a.[15] The utility function $u(\cdot)$ is twice differentiable, unbounded, increasing, and concave (i.e., more money is preferred to less and agents are risk averse). As before, the firm hires agents by making take-it-or-leave-it offers. Let $\bar{u}$ be an agent's reservation utility. To eliminate *exogenous* fixed costs, assume that

$$u^{-1}(\bar{u}) = 0. \tag{6}$$

Unlike the previous sections, assume now that an agent's action is unobservable by the firm, so contracts cannot be written contingent upon it. If they could be, then the firm's cost of effort would be $u^{-1}(a)$, which, given (6), would mean the firm would hire an arbitrarily large number of agents each of whom would supply an arbitrarily small amount of effort.

Assume, instead, that there are two verifiable states: the firm does well and the firm does poorly. Let the probability that the firm does well be $P\left(\sum_{n=1}^{N} a_n\right)$, where $P(\cdot)$ is a twice-differentiable, increasing, and concave function. Assume, moreover that $P(0) = 0$ and $P(A) < 1$ for all A.

[14] Note this is the expected cost whether he is paid piece rate (i.e., his actual compensation is $c(t) + \bar{w}$, where t is the realized number of problems he faces) or whether he is paid prospectively (i.e., his actual compensation is $\mathcal{C}_m(T)$).

[15] For a *single-agent* model, there is no loss in generality from parameterizing the disutility of action function as the identity function, as done here (see Rogerson, 1985). For a multi-agent model, such as the one explored here, this does represent a loss of generality because of what it implies about the additivity of action across agents. This is not, however, a serious problem—the main points of the model would still arise with a more general parameterization.

A contract for agent n is a pair of contingent payments, (w_1^n, w_2^n), where w_1^n is the payment if the firm does poorly and w_2^n is the payment if the firm does well.

Suppose that the firm would like to induce an agent to choose a_n. Let the equilibrium actions of the other agents be $\{\hat{a}_j\}_{j \neq n}$. The contract (w_1^n, w_2^n) will induce the agent to choose a_n if a_n is his best response to the contract; that is, if

$$a_n = \arg\max_a P\left(a + \sum_{j \neq n} \hat{a}_j\right) u\left(w_2^n\right) + \left[1 - P\left(a + \sum_{j \neq n} \hat{a}_j\right)\right] u\left(w_1^n\right) - a. \tag{7}$$

Since $P(\cdot)$ is concave, (7) can be replaced by the first-order condition:

$$P'\left(a_n + \sum_{j \neq n} \hat{a}_j\right)\left(u\left(w_2^n\right) - u\left(w_1^n\right)\right) - 1 = 0. \tag{8}$$

The agent will accept the contract (w_1^n, w_2^n) provided his expected utility under the contract is at least as great as his reservation utility:

$$P\left(a_n + \sum_{j \neq n} \hat{a}_j\right) u\left(w_2^n\right) + \left[1 - P\left(a_n + \sum_{j \neq n} \hat{a}_j\right)\right] u\left(w_1^n\right) - a_n \geq \bar{u}. \tag{9}$$

The firm wants to offer a contract that induces the desired action, is acceptable to the agent, and minimizes its expected wage bill. Solving this problem yields

Lemma 4 *Conditional on the equilibrium actions of the other agents, $\{\hat{a}_j\}_{j \neq n}$, the optimal contract for inducing a_n, $a_n > 0$, from agent n is*

$$\begin{aligned} w_1^n &= u^{-1}\left(\bar{u} + a_n - \frac{P\left(a_n + \sum_{j \neq n} \hat{a}_j\right)}{P'\left(a_n + \sum_{j \neq n} \hat{a}_j\right)}\right) \text{ and} \\ w_2^n &= u^{-1}\left(\bar{u} + a_n + \frac{1 - P\left(a_n + \sum_{j \neq n} \hat{a}_j\right)}{P'\left(a_n + \sum_{j \neq n} \hat{a}_j\right)}\right). \end{aligned} \tag{10}$$

Like the full-information case, the firm minimizes its expected cost of inducing *total* effort A by inducing each of its N agents to expend effort A/N:

Lemma 5 *Suppose the firm wishes to induce total effort A. Then it is cost-minimizing for it to induce each of its N agents to expend effort A/N. That is, it minimizes cost by offering the contract*

$$w_1 = u^{-1}\left(\bar{u} + \frac{A}{N} - \frac{P(A)}{P'(A)}\right) \text{ and } w_2 = u^{-1}\left(\bar{u} + \frac{A}{N} + \frac{1 - P(A)}{P'(A)}\right)$$

to each agent.

Value of A	Optimal Firm Size
$(0, 1.29)$	1
$(1.29, 2.39)$	2
$(2.39, 3.43)$	3
$(3.43, 4.45)$	4
$(4.45, 5.47)$	5
$(5.47, 6.47)$	6
$(6.47, 7.48)$	7
$(7.48, 8.48)$	8
$(8.48, 9.48)$	9

Table 1: Optimal Firm Size as a Function of A

In light of Lemmas 4 and 5, the expected cost of inducing *total* effort A from N agents is

$$\mathcal{C}_N(A) = NP(A)\,u^{-1}\left(\bar{u} + \frac{A}{N} + \frac{1-P(A)}{P'(A)}\right) + N(1-P(A))\,u^{-1}\left(\bar{u} + \frac{A}{N} - \frac{P(A)}{P'(A)}\right).$$

Note that since $P(0) = 0$, this definition of $\mathcal{C}_N(A)$ is valid for all $A \geq 0$. It is readily verifiable that $\mathcal{C}_N(\cdot)$ is twice differentiable on $(0, \infty)$ and that $\mathcal{C}_N(0) = 0$ for all N.

The goal here is to show that the curves $\mathcal{C}_N(\cdot)$ cross in a manner similar to that in Figure 1. To this end, consider

Proposition 8 *Let $M > N$, then there exists an $\tilde{A} > 0$ such that $\mathcal{C}_N(A) < \mathcal{C}_M(A)$ for all $A \in \left(0, \tilde{A}\right)$.*

Proposition 8 establishes that if a firm wished to induce a small action in aggregate, then it does better to employ few agents rather than many agents. What Proposition 8 unfortunately does not do is establish that $\tilde{A}$ is finite; that is, that the cost curves eventually cross. Without narrowing the properties of $P(\cdot)$ or $u(\cdot)$ further, such a result cannot be established. On the other hand, it is straightforward to find examples in which the curves do eventually cross.

Example 2 *Let*

$$P(A) = 1 - \frac{1}{A+1},\; u(w) = \ln(w+1)\,,\; \text{and } \bar{u} = 0.$$

Then it can be shown that for each N there is an interval of A's such that $\arg\min_K \mathcal{C}_K(A) = N$ *for A's in this interval.*[16] *Table 1 provides data for $N = 1, \ldots, 9$.*

To understand the crossing pattern of the cost curves, note that each wage term has the form

$$u^{-1}\left(\frac{A}{N} + \xi(A)\right);$$

that is, consists of a term that decreases with N and a term that is independent of N. The term that is decreasing with N combined with the convexity of $u^{-1}(\cdot)$ creates the returns-to-scale effect that makes employing multiple agents preferable to employing only a few agents. The term that is

[16] The program for verifying this is available from the author upon request.

independent of N is essentially a fixed cost of agency, which makes employing few agents preferable to employing many. When the returns-to-scale effect is small—i.e., when A is small—the fixed-cost effect dominates, so few agents are better than many. When the returns-to-scale effect is large—i.e., when A is large—the returns-to-scale effect can dominate, so many agents are better.

A fixed cost of agency exists because each agent must be on the same margin with respect to his choice of effort. A margin, moreover, that is fixed by the *total* effort to be supplied rather than by the individual effort to be supplied. In this sense, the problem with multiple agents is similar to the teams problem considered by Holmstrom (1982). In that paper, individual rewards for effort are based on total production. Consequently, to give each individual the correct incentives, each individual must receive 100% of the *total* marginal increase in production. Here, each agent must be given 100% of the bonus—$(1 - P(A))/P'(A)$—for doing well and 100% of the penalty—$P(A)/P'(A)$—for doing poorly. As usual in an agency problem, providing incentives is costly; hence, there is a fixed cost of incentives for each agent employed.

These "fixed costs," like the fixed costs considered in earlier sections, can serve as the explanation for many aspects of organizational structure. Moreover, agency problems can generate other non-convexities that act like the gaps considered in Section 2: Even if $P(\cdot)$ is concave, $\mathcal{C}_N(\cdot)$ need not be convex, which also creates gaps—see Hermalin (1994) for examples and further discussion.

5 Conclusion

This paper has sought to show how fixed costs help to determine organizational structure; similar to how they help to determine industry structure.

In Section 2, it was shown that fixed costs can generate regions of non-convexity or "gaps" in cost functions. These gaps mean, first, that firm behavior can exhibit discrete jumps in response to small changes in the firm's environment. This in turn offers explanations for phenomena such as layoffs and downsizing.

These gaps also introduce non-convexities into games played among firms. These non-convexities can result in asymmetric equilibria among the firms; indeed, they can be *necessary* for asymmetric equilibria to exist. Fixed costs can, thus, help with the important research question of why are firms different (see Rumelt *et al.*, 1994, pages 225–228, for a discussion of this question's importance).

In Section 3, fixed costs were used to explain the pyramid shape common to most hierarchies. The idea was that because superiors solve the problems left unsolved by their subordinates, higher layers of the hierarchy have, in expectation, less work to do. To economize on fixed costs, this less work is spread over fewer employees. Consequently, there are fewer employees at higher layers than at lower layers—the hierarchy is pyramid shaped.

Finally, Section 4 showed that these fixed costs need not be assumed exogenously, but could instead arise endogenously because of agency problems within the organization.

Two final points are worth making. First, it is *not* the goal of this paper to argue that fixed-cost explanations of organizational structure are superior to other explanations or that they should supplant them.[17] Rather, the goal is to offer a complementary perspective on organizational structure, which has the benefit of providing a unifying view of many seemingly disparate organizational phenomena (e.g., downsizing and pyramidal hierarchies).

[17] For instance, communication-cost-based explanations (e.g., Bolton and Dewatripont, 1994, Segal, 1996) also offer important insights.

Second, there is clearly much work to be done. The models presented here are, of necessity, extremely simple. Certainly, they can be extended. Moreover, there are many other organizational phenomena that can be analyzed through the lens of fixed costs.

Appendix

Proof of Lemma 1: Clearly, $C(\cdot;N)$ is continuous. If $M > N$, then $C(0;M) = M\bar{w} > N\bar{w} = C(0;N)$. Define a^* to be the solution to

$$\bar{w} + c(a) - ac'(a) = 0$$

(the proof of Proposition 2 establishes that a^* exists). Consider the minimization program

$$\min_K C(Ma^*;K).$$

The first-order condition is

$$\bar{w} + c\left(\frac{Ma^*}{K}\right) - \frac{Ma^*}{K}c'\left(\frac{Ma^*}{K}\right) = 0.$$

By definition of a^*, the solution is $K = M$. Hence, $C(Ma^*;N) > C(Ma^*;M)$. The functions $C(\cdot;N)$ and $C(\cdot;M)$ cross, at least one time from above. Finally,

$$\frac{\partial C}{\partial N \partial A} = -c''\left(\frac{A}{N}\right)\frac{A}{N^2} < 0, \tag{11}$$

which means that the marginal cost of A is less, the greater is N. Hence the functions $C(\cdot;N)$ and $C(\cdot;M)$ cross at most once. ■

Proof of Proposition 1: Recall that $C(\cdot;N)$ crosses $C(\cdot;N+1)$ from below. Suppose, first, that A_N did maximize profit. Then necessary conditions for A_N to satisfy are

$$\begin{aligned} \beta'(A_N) - \frac{\partial C(A_N;N)}{\partial A} &\geq 0 \text{ and} \\ \beta'(A_N) - \frac{\partial C(A_N;N+1)}{\partial A} &\leq 0. \end{aligned}$$

Rearranging and simplifying, these yield

$$\frac{\partial C(A_N;N+1)}{\partial A} \geq \frac{\partial C(A_N;N)}{\partial A}. \tag{12}$$

But the marginal cost of action is *falling* in N (see (11), which is inconsistent with (12). By contradiction, A_N cannot maximize profits. The result then follows by continuity. ■

Proof of Lemma 2: Let $\hat{A}$ be a solution to the **-program. There are two cases to consider: $C^*\left(\hat{A}\right) = C^{**}\left(\hat{A}\right)$ and $C^*\left(\hat{A}\right) > C^{**}\left(\hat{A}\right)$ (by definition, $C^*\left(\hat{A}\right) \geq C^{**}\left(\hat{A}\right)$). Suppose $C^*\left(\hat{A}\right) = C^{**}\left(\hat{A}\right)$, then

$$\begin{aligned} \beta\left(\hat{A}\right) - C^*\left(\hat{A}\right) &= \beta\left(\hat{A}\right) - C^{**}\left(\hat{A}\right) \\ &\geq \beta(A) - C^{**}(A) \text{ for all } A \in \left[0,\bar{A}\right] \text{ (def'n of maximum)} \\ &\geq \beta(A) - C^*(A) \text{ for all } A \in \left[0,\bar{A}\right]; \end{aligned}$$

so $\hat{A}$ is the solution to the *-program. Suppose $C^{*}\left(\hat{A}\right) > C^{**}\left(\hat{A}\right)$. Define

$$\begin{aligned} A^{+} &= \min\left\{A|A > \hat{A} \text{ and } C^{*}(A) = C^{**}(A)\right\} \text{ and} \\ A^{-} &= \max\left\{A|A < \hat{A} \text{ and } C^{*}(A) = C^{**}(A)\right\} \end{aligned}$$

(because $C^{*}(0) = C^{**}(0)$ and $C^{*}\left(\bar{A}\right) = C^{**}\left(\bar{A}\right)$ by construction and $C^{*}(\cdot)$ is continuous, A^{+} and A^{-} are well defined). $C^{**}(A)$ is a convex combination of $C^{**}(A^{-})$ and $C^{**}(A^{+})$ for $A \in (A^{-}, A^{+})$, including $\hat{A}$. Given that $\beta(\cdot)$ is affine, it follows that $\beta(A^{+}) - C^{**}(A^{+}) = \beta\left(\hat{A}\right) - C^{**}\left(\hat{A}\right)$. Hence, $\beta(A^{+}) - C^{*}(A^{+}) > \beta\left(\hat{A}\right) - C^{*}\left(\hat{A}\right)$: $\hat{A}$ is not a solution to the *-program.

Turning to the "moreover" part, let $\hat{A}$ be a solution to the *-program. Suppose $C^{*}\left(\hat{A}\right) > C^{**}\left(\hat{A}\right)$. Define A^{-} and A^{+} as before. Because $C^{**}\left(\hat{A}\right)$ is a convex combination of $C^{**}(A^{-})$ and $C^{**}(A^{+})$ and because $\beta(\cdot)$ is affine,

$$\max\left\{\beta\left(A^{+}\right) - C^{**}\left(A^{+}\right), \beta\left(A^{-}\right) - C^{**}\left(A^{-}\right)\right\} \geq \beta\left(\hat{A}\right) - C^{**}\left(\hat{A}\right).$$

But, then,

$$\begin{aligned} \max\left\{\beta\left(A^{+}\right) - C^{*}\left(A^{+}\right), \beta\left(A^{-}\right) - C^{*}\left(A^{-}\right)\right\} &\geq \beta\left(\hat{A}\right) - C^{**}\left(\hat{A}\right) \\ &> \beta\left(\hat{A}\right) - C^{*}\left(\hat{A}\right); \end{aligned}$$

which contradicts $\hat{A}$ being a solution to the *-program. ■

Proof of Proposition 2: First, let's see that a^{*} is defined: Since $\bar{w} + c(a) > 0$ all a and $c(\cdot)$ is convex, there must be a line tangent to $\bar{w} + c(a)$ that passes through the origin. But this, then, implies (2) has a solution; that is, a^{*} is well defined.

Let $\bar{A}$ equal some $A \geq \bar{N}a^{*}$. By construction $C^{**}(A) \leq C^{*}(A)$ and $C^{**}(\cdot)$ is convex, which means $C^{**}(\cdot)$ is never above the line segment connecting two points on $C^{*}(\cdot)$. Specifically, let $l(A)$ be the line passing through $(A_L, C^{*}(A_L))$and $(A_H, C^{*}(A_H))$, where $A_L < A_H$. Then $C^{**}(A) \leq l(A)$ for $A \in (A_L, A_H)$.

Claim: $C(A; N) \geq Ac'(a^{*})$ *for all* A *and* N *and equal* only *at* $A = \hat{A}_N$.

Proof of Claim: Solving the program

$$\min_{A} C(A; N) - Ac'(a^{*})$$

yields the first-order condition

$$c'\left(\frac{A}{N}\right) - c'(a^{*}) = 0.$$

The solution to this program is $A = \hat{A}_N$. The difference

$$\begin{aligned} C\left(\hat{A}_N; N\right) - \hat{A}_N \cdot c'(a^{*}) &= N\bar{w} + Nc\left(\frac{\hat{A}_N}{N}\right) - \hat{A}_N \cdot c'(a^{*}) \\ &= N \cdot (\bar{w} + c(a^{*})) - Na^{*}c'(a^{*}) \\ &= 0, \end{aligned}$$

where the last equality follows from (2). Since $C(\cdot;N)$ is strictly convex, the result follows.□

From the claim, $C^*(A) \geq Ac'(a^*)$ for all A and equal *only* at the points $A = \hat{A}_N$. Let $l(A)$ be the line passing through $(0,0)$ and $\left(\hat{A}_{\bar{N}}, C^*\left(\hat{A}_{\bar{N}}\right)\right)$. By construction, $l(A) = Ac'(a^*)$. It has, thus, been shown that

$$C^{**}(A) \leq l(A) \leq C^*(A) \text{ for all } A \in \left[\hat{A}_1, \hat{A}_{\bar{N}}\right].$$

Moreover, because the second inequality holds strictly for all $A \notin \left\{0, \hat{A}_1, \ldots, \hat{A}_{\bar{N}}\right\}$, it follows from Lemma 2 that all $A \in \left(0, \hat{A}_{\bar{N}}\right] - \left\{\hat{A}_1, \ldots, \hat{A}_{\bar{N}}\right\}$ are in the gap. Finally, it is straightforward to show that

$$C^{**}(A) = C^*(A) = C\left(A; \bar{N}\right) \text{ for } A \in \left[\hat{A}_{\bar{N}}, \bar{A}\right],$$

which completes the proof. ■

Proof of Proposition 3: Define

$$A_N^* = \arg\max_A \theta A + k - C(A;N).$$

Then

$$\theta = c'\left(\frac{A_N^*}{N}\right).$$

Define a_θ^* by $\theta = c'(a_\theta^*)$. Then it follows for all N that

$$\max_A \theta A + k - C(A;N) = \theta N a_\theta^* + k - N\bar{w} - Nc(a_\theta^*).$$

Consider the program

$$\max_N \theta N a_\theta^* + k - N\bar{w} - Nc(a_\theta^*). \tag{13}$$

This is affine in N. Provided $\theta a_\theta^* \neq \bar{w} + c(a_\theta^*)$, this means that the solution is

$$\begin{aligned} 0 \text{ if } \theta a_\theta^* &< \bar{w} + c(a_\theta^*), & (14)\\ \text{All } N \text{ if } \theta a_\theta^* &= \bar{w} + c(a_\theta^*), \text{ and} & (15)\\ \bar{N} \text{ if } \theta a_\theta^* &> \bar{w} + c(a_\theta^*). & (16) \end{aligned}$$

Since $c(\cdot)$ is convex, a^* is unique, which means (15) holds only if $\theta = c'(a^*)$. Moreover, its convexity implies that if $\theta >$ (alt. $<$) $c'(a^*)$, then $a_\theta^* >$ (alt. $<$) a^*. The line tangent to $\bar{w} + c(a)$ at a^* has a negative (alt. positive) intercept if $a_\theta^* >$ (alt. $<$) a^*, from which it follows that (16) holds if $\theta > c'(a^*)$ and (14) holds if $\theta < c'(a^*)$.

Let N^* be the (a) solution to (13). Define

$$l(A) = \theta A + k^*,$$

where

$$k^* = N^* \cdot (\bar{w} + c(a_\theta^*) - \theta a_\theta^*).$$

By construction, $C(A_{N^*}^*; N^*) = l(A_{N^*}^*)$. All that is left to show is that $C^*(A) \geq l(A)$, since, then,

$$\begin{aligned} A_{N^*}^* &= \arg\max_A \theta A + k - C^*(A) \text{ and} & (17)\\ N^* &= \arg\min_N C(A_{N^*}^*; N); \end{aligned}$$

that is, N^* would, indeed, be the optimal size. Consider the program

$$\min_A C(A;N) - l(A)$$

for all N. The first-order condition is

$$c'\left(\frac{A}{N}\right) - \theta = 0.$$

Hence, $C(A;N) - l(A) \geq C(Na_\theta^*;N) - l(Na_\theta^*)$ for all A and equal only if $A = A_N^*$.

$$\begin{aligned} C(Na_\theta^*;N) - l(Na_\theta^*) &= N\bar{w} + Nc(a_\theta^*) - N\theta a_\theta^* - k^* \\ &= N(\bar{w} + c(a_\theta^*) - \theta a^*) - N^*(\bar{w} + c(a_\theta^*) - \theta a^*) \\ &\geq 0 \text{ (by def'n of } N^*) \end{aligned}$$

and equal only if $N = N^*$.

Finally, if $N^* = 0$, the firm would be shutdown—benefit would be zero (benefit, recall, is discontinuous at $A = 0$). If $k > \bar{w}$—i.e., $\underline{N}(k) = 1$—this would be sub-optimal. Hence, when $\underline{N}(k) = 1$, the firm will have one employee. ■

Proof of Proposition 4: Firm i's equilibrium choice of A_i must satisfy[18]

$$I - S(A_i + A_j) - SA_i - C^{*\prime}(A_i) \begin{cases} = 0 \text{ if } A_i > 0 \\ \leq 0 \text{ if } A_i = 0 \end{cases}, \tag{18}$$

where A_j is the rival's equilibrium choice. Suppose $A_1 > A_2$. Then

$$I - S(A_1 + A_2) - SA_1 < I - S(A_1 + A_2) - SA_2.$$

But since $C^*(\cdot)$ is convex, this expression combined with (18) implies the contradiction $A_1 < A_2$. Since the indices 1 and 2 are arbitrary, all pure-strategy equilibria must be symmetric. Rewriting (18) in this light, yields

$$I - 3SA - C^{*\prime}(A) \begin{cases} = 0 \text{ if } A > 0 \\ \leq 0 \text{ if } A = 0 \end{cases}.$$

The first term is decreasing in A while the second term is increasing in A, so they can intersect at most once for $A \geq 0$. This establishes uniqueness. ■

Proof of Proposition 5: Here $\theta_i = \frac{1}{\phi(A_j)}$. It is readily shown that the result then follows as a corollary of Proposition 3. ■

Proof of Proposition 6: Suppose there were a last employee for a given problem. Then if it were optimal to have him, but no more employees, then

$$\max_\alpha b_t\pi_t(\alpha) - c(\alpha) \geq \max_{\alpha_1,\alpha_2} b_t\pi_t(\alpha_1) + (1 - \pi_t(\alpha_1))\left[b_t\pi_t(\alpha_2) - c(\alpha_2)\right] - c(\alpha_1).$$

Using stars to denote the solutions to the two maximization programs, it is clear that $\alpha^* = \alpha_2^*$. So this last inequality can be rewritten as

$$0 \geq \max_{\alpha_1} b_t\pi_t(\alpha_1) - c(\alpha_1) - \pi_t(\alpha_1)\left[b_t\pi_t(\alpha^*) - c(\alpha^*)\right].$$

[18] Since $C^*(\cdot)$ is convex, it is differentiable almost everywhere. To keep the proof short, but without loss of generality, I will simply treat it as differentiable everywhere.

But this can't be—set $\alpha_1 = \alpha^*$, then, since $\pi_t(\alpha^*) < 1$, the right-hand side is strictly positive. Hence, if it pays to employ one employee, it pays to employ—potentially—an infinite number of them.

Since an infinite number of employees will, potentially, deal with a problem and the probability of solving the problem is independent of past failures, it will be optimal for each employ to expend the same effort. Call it, α^{**}. Expected profits are, then,

$$\sum_{t=1}^{\infty} \pi(\alpha^{**})(1-\pi(\alpha^{**}))^{t-1}(b - tc(\alpha^{**})).$$

Simplifying (exploiting the fact that t is distributed geometrically), expected profit is

$$b - \frac{c(\alpha^{**})}{\pi(\alpha^{**})}.$$

■

Proof of Lemma 3:

$$\begin{aligned}
\frac{d\mathcal{C}_m(T)}{dm} &= \frac{d\hat{q}_m}{dm} \times \frac{1}{\hat{q}_m(1-\hat{q}_m)} \sum_{t=0}^{T} \binom{T}{t} \hat{q}_m^t (1-\hat{q}_m)^{T-t} c(t)(t - \hat{q}_m T) \\
&= (1-\hat{\pi})^{m-1} \ln(1-\hat{\pi}) \times \frac{1}{\hat{q}_m(1-\hat{q}_m)} \times \operatorname{cov}(c(t), t).
\end{aligned}$$

Since the covariance of $c(t)$ and t is positive ($c(t)$ is increasing), $\operatorname{cov}(c(t), t) > 0$. Of course, $\ln(1-\hat{\pi}) < 0$, so it follows that $d\mathcal{C}_m(T)/dm < 0$. ■

Proof of Proposition 7: Considering merging two positions, one which has a maximum of T problems and another that has a maximum of S problems, into one position with a maximum of $S+T$ problems. Define

$$p_{m,T}(t) = \binom{T}{t} (1-\hat{q}_m)^{T-t} \hat{q}_m^t.$$

Note that

$$\frac{dp_{m,T}(t)}{dm} = \frac{d\hat{q}_m}{dm} p_{m,T}(t) \frac{t - \mathbb{E}t}{\hat{q}_m(1-\hat{q}_m)}.$$

Then the firm's expected cost if it leaves the positions unmerged is

$$\begin{aligned}
\mathcal{U} &= 2\bar{w} + \sum_{t=0}^{T} p_{m,T}(t) c(t) + \sum_{s=0}^{S} p_{m,S}(s) c(s) \\
&= 2\bar{w} + \sum_{t=0}^{T} p_{m,T}(t) \sum_{s=0}^{S} p_{m,S}(s)(c(t) + c(s)).
\end{aligned}$$

Its expected cost if it merges the positions is

$$\mathcal{M} = \bar{w} + \sum_{t=0}^{T} p_{m,T}(t) \sum_{s=0}^{S} p_{m,S}(s) c(s+t).$$

Hence,

$$\mathcal{U}-\mathcal{M}=\bar{w}+\sum_{t=0}^{T}p_{m,T}\left(t\right)\sum_{s=0}^{S}p_{m,S}\left(s\right)\left[c\left(s\right)+c\left(t\right)-c\left(s+t\right)\right].$$

Consider:

$$\begin{aligned}\frac{d\left(\mathcal{U}-\mathcal{M}\right)}{dm} &= \frac{d\hat{q}_m}{dm}\frac{1}{\hat{q}_m\left(1-\hat{q}_m\right)}\times\left[\sum_{s=0}^{S}p_{m,S}\left(s\right)\sum_{t=0}^{T}p_{m,T}\left(t\right)\left(t-\mathbb{E}t\right)\left(c\left(t\right)-c\left(s+t\right)\right)\right. \\ &\quad \left.+\sum_{t=0}^{T}p_{m,T}\left(t\right)\sum_{s=0}^{S}p_{m,S}\left(s\right)\left(s-\mathbb{E}s\right)\left(c\left(s\right)-c\left(s+t\right)\right)\right] \\ &= \frac{d\hat{q}_m}{dm}\frac{1}{\hat{q}_m\left(1-\hat{q}_m\right)}\times\left[\mathbb{E}_s\left\{\operatorname{cov}\left(t,c\left(t\right)-c\left(s+t\right)\right)\right\}+\mathbb{E}_t\left\{\operatorname{cov}\left(s,c\left(s\right)-c\left(s+t\right)\right)\right\}\right].\end{aligned}$$

Since $c\left(\cdot\right)$ is convex, $c\left(t\right)-c\left(s+t\right)$ is *decreasing* in t. Hence, $\operatorname{cov}\left(t,c\left(t\right)-c\left(s+t\right)\right)<0$ for all s. Similarly, $\operatorname{cov}\left(s,c\left(s\right)-c\left(s+t\right)\right)<0$ for all t. Since $d\hat{q}_m/dm<0$, it follows that $d\left(\mathcal{U}-\mathcal{M}\right)/dm>0$. ■

Proof of Lemma 4: Grossman and Hart (1983) showed that (9) is binding under the optimal contract. So the firm has a minimization problem in two variables, w_1^n and w_2^n, with two binding constraints, (9) and (8). The solution to this problem can, therefore, be found by solving the constraints. One can readily verify that (10) is the solution to these two constraints. ■

Proof of Lemma 5: If $A=0$, then clearly each agent must be induced to choose 0. So suppose $A>0$. From Lemma 4, the firm's problem is

$$\begin{aligned}&\min_{\{a_n\}}P\left(A\right)\sum_{n=1}^{N}u^{-1}\left(\bar{u}+a_n+\frac{1-P\left(A\right)}{P'\left(A\right)}\right)+\left(1-P\left(A\right)\right)\sum_{n=1}^{N}u^{-1}\left(\bar{u}+a_n-\frac{P\left(A\right)}{P'\left(A\right)}\right)\\ &\text{subject to }\sum_{n=1}^{N}a_n=A.\end{aligned}$$

The result follows from the convexity of $u^{-1}\left(\cdot\right)$. ■

Proof of Proposition 8: The right derivative of $\mathcal{C}_N\left(A\right)$ evaluated $A=0$ is

$$NP'\left(0\right)u^{-1}\left(\bar{u}+\frac{1}{P'\left(0\right)}\right)+N\left(\frac{1}{N}-1\right)\left(u^{-1}\right)'\left(\bar{u}\right).$$

Differentiating this expression with respect to N yields

$$P'\left(0\right)u^{-1}\left(\bar{u}+\frac{1}{P'\left(0\right)}\right)-\left(u^{-1}\right)'\left(\bar{u}\right). \tag{19}$$

Expression (19) is positive, since the convexity of $u^{-1}\left(\cdot\right)$ implies the following about the first-order Taylor series expansion:

$$u^{-1}\left(\bar{u}+\frac{1}{P'\left(0\right)}\right)-\left(u^{-1}\right)'\left(\bar{u}\right)\frac{1}{P'\left(0\right)}>0.$$

Since (i) the right derivative of $\mathcal{C}_M\left(\cdot\right)$ is greater than the right of derivative of $\mathcal{C}_N\left(\cdot\right)$ at zero; (ii) $\mathcal{C}_M\left(0\right)=\mathcal{C}_N\left(0\right)$; and (iii) $\mathcal{C}_M\left(\cdot\right)$ and $\mathcal{C}_N\left(\cdot\right)$ are continuous; the proposition follows. ■

References

Aron, D. J. (1988). Ability, moral hazard, firm size, and diversification. *RAND Journal of Economics*, **19**, 72–87.

Bolton, P. and Dewatripont, M. (1994). The firm as a communication network. *Quarterly Journal of Economics*, **109**, 809–952.

Calvo, G. and Wellisz, S. (1978). Supervision, loss of control, and the optimum size of the firm. *Journal of Political Economy*, **86**, 943–952.

Geanakoplos, J. and Milgrom, P. (1991). A theory of hierarchies based on limited managerial attention. *Journal of the Japanese and International Economies*, **5**, 205–225.

Grossman, S. and Hart, O. (1983). An analysis of the principal-agent problem. *Econometrica*, **51**, 7–45.

Hermalin, B. E. (1994). Heterogeneity in organizational form: Why otherwise identical firms choose different incentives for their managers. *RAND Journal of Economics*, **25**, 518–537.

Holmstrom, B. (1982). Moral hazard in teams. *Bell Journal of Economics*, **13**, 324–340.

Lucas, R. (1978). On the size distribution of business firms. *Bell Journal of Economics*, **9**, 508–523.

McAfee, R. P. and McMillan, J. (1995). Organizational diseconomies of scale. *Journal of Economics & Management Strategy*, **4**, 399–426.

New York Times (1996). *The Downsizing of America*. Times Books, Random House, New York.

Rogerson, W. (1985). The first-order approach to principal-agent problems. *Econometrica*, **53**, 1357–1368.

Rumelt, R. P., Schendel, D. E., and Teece, D. J., editors (1994). *Fundamental Issues in Strategy: A Research Agenda*. Harvard Business School Press, Boston.

Sah, R. K. and Stiglitz, J. E. (1986). The architecture of economic systems: Hierarchies and polyarchies. *American Economic Review*, **76**, 716–727.

Segal, I. R. (1996). Modeling the managerial task. Working Paper, Department of Economics, University of California at Berkeley.

van Tiel, J. (1984). *Convex Analysis*. John Wiley and Sons, New York.

Williamson, O. (1967). Hierarchical control and the optimum size of the firm. *Journal of Political Economy*, **75**, 123–138.