Optimal Coordination in Hierarchies

Andrea Patacconi
University of Oxford

1 August 2005

Abstract

This paper studies the optimal allocation of coordination responsibilities in organizations where duplication of effort is a serious concern. The planner's objective is to minimize a weighted average of the wage bill and the cost of delay. The paper provides conditions under which, in balanced hierarchies, communication effort is increasing and the span of control is decreasing as one travels up the hierarchy, with equalities holding if wages are negligible relative to the weight attached to the cost of delay. The analysis suggests that concerns for fast decision-making may be key in explaining the recent trend towards empowerment in firms. Several variants of the basic model are studied, including one focusing on communicative skills and another in which, as urgency increases, the optimal span of control increases and the hierarchy flattens. Evidence supporting these results is discussed.

Keywords: Coordination, Hierarchy, Duplication, Delay, Information Processing.
JEL Classification: D21, L23

[1]Postal address: Andrea Patacconi, Department of Economics, Manor Road Building, Manor Road, Oxford, OX1 3UQ, United Kingdom. E-mail: andrea.patacconi@economics.ox.ac.uk.

I am particularly indebted to Meg Meyer for her guidance at every stage of the project and to John Quah for many comments that have greatly improved the paper. I also would like to thank Dan Anderberg, Ernesto Dal Bo, Wouter Dessein, Erik Eyster, Niko Matouschek John Roberts, Armin Schmutzler, seminar participants at the University of Oxford, Royal Holloway and the 2004 Congress of the European Economic Association at the Universitat Carlos III in Madrid for helpful conversations and remarks. Financial support from the Economics Department at the University of Oxford and the Royal Economic Society is gratefully acknowledged. All remaining errors are mine.

1 Introduction

The division of labor requires the existence of means by which specialized activities can be properly coordinated. Within the firm, for instance, decisions are usually taken through a complex system of authority levels which is to a large extent a matter of planning. Indeed, one of the fundamental tasks of the entrepreneur (or the management in large organizations) is to ensure that the activities within the firm are properly organized and coordinated.

This paper is concerned with the optimal organization of coordination activities within the firm – identified here as a hierarchy. Besanko et al. (2000) argue that "coordination involves the flow of information to facilitate subunit decisions that are consistent with each other and with organizational objectives" (p.549). This paper deals with a specific but important aspect of that definition, namely the efficient division of labor within hierarchical organizations. A key assumption I put forward is that by spending more effort on coordination and planning activities such as work scheduling, arranging meetings or the exchanging of information with colleagues, managers can reduce overlaps among tasks and therefore wasteful duplications of effort.[1]

Duplication of effort has long been recognized as a serious concern, especially in large organizations and the public sector (Simon et al. (1950)). Chandler (1990), for instance, reports that, by the end of 1920, the director of Du Pont's Chemical Department, together with the departmental research directors, took on the responsibility of coordinating the different manufacturing departments "so that overlapping of the research programs [...] may be avoided as completely as possible" (p.182). Similarly, Stuart Rice, in his analysis of the role and management of the Federal Statistical System, rhetorically asks why "if I head an electric utility, should I be asked to report my total pay-roll or number of employees, or both, to six different federal offices, on 12 different forms, at intervals varying from one month to one year?" (1940, p.482). He emphasizes, as this paper does, that duplication of effort usually stems from the decentralization of agencies (i.e., the division of labor) and that coherence in organization can be attained only by an item to item adjustment of each task and process to every other related task and process, that is, by coordinating interrelated activities.

The recent wave of corporate investments in information technology is also in part motivated by the need to reduce duplications of effort. In the pharmaceutical industry, for instance, companies test huge numbers of different compounds in search for new drugs, and information systems such as Intranets are routinely used to share information, improve collaboration among organization units and minimize duplications of research effort. Two examples among many are SmithKline Beecham, which runs virtual

[1] Of course in reality not all duplications need to be wasteful. Many authors, for instance, have argued that redundancies may actually be an optimal response to the possibility of 'misunderstandings' in the communication process (e.g., Chwe (1995)). However, in my final remarks, I will argue that the present model can be reinterpreted along these lines.

libraries that can be consulted by each of its employees, and Glaxo-Wellcome, whose research scientists can access complex chemical information including structural images on their internal Intranet (see www.skyrme.com/insights/25intra.htm for more examples from different industries).

Notwithstanding its practical relevance, it is fair to say that the problem of duplication of effort has received little attention in the economics literature.[2] Previous work, in fact, has focused on other aspects of 'loss of control' in organizations, most notably the loss of useful information (e.g., Williamson (1967) and Vayanos (2003)) and shirking (e.g., Calvo and Wellisz (1978,1979) and Qian (1994)). This paper extends this literature by exploring the implications of a tradeoff between duplication of effort and coordination costs in hierarchical organizations that must carry out a task of given size. The organization's task can be interpreted either as producing goods such as cars or drugs or as processing information (see below). In line with previous work, upper-level managers can delegate their (sub)tasks to a few subordinates, whom I call a workgroup, in order to reduce delay. However, in this model, superiors must also spend time and effort coordinating their subordinates if they want to reduce wasteful duplications of tasks and effort.[3] In particular, I assume that the percentage of a delegated task which is duplicated, captured by the function $D(n, c)$, depends on both n, the number of subordinates a manager has (that is, his span of control) and the superior's communication effort c (also called coordination effort). Thus, $1/D(n, c)$ can be interpreted as a measure of coordination at a given layer. Attention in the model will be restricted to balanced hierarchies, that is, hierarchies where all the managers at the same level choose the same communication effort and span of control.[4] A hierarchy can therefore be fully characterized by a triple: a vector of communication efforts, a vector of spans of control, and the number of levels in the hierarchy. Importantly, since communication effort and span of control will be decision variables at each level of the hierarchy, this will allow me to compare their optimal levels across layers.

Two types of balanced hierarchies are considered in this paper, "production hierarchies" and "information processing hierarchies". In the basic model, a neat distinction is drawn between bottom-level employees (workers), who are engaged in production, and the managers at higher levels, whose only job is to coordinate their direct subordinates. In this scenario the task is best interpreted as the production of goods, so the organization is termed a production hierarchy. Many well-known models share this basic structure, including Calvo and Wellisz (1979), Keren and Levhari (1979) and Qian (1994).

In information processing hierarchies, by contrast, performing the task requires the top manager to acquire some relevant information. As in the production scenario, upper-level managers are in charge of coordinating the work of their subordinates and raw information is processed only by bottom-level man-

[2] A notable exception is Bolton and Farell (1990); their focus however is not on the internal organization of firms.

[3] Alternatively, subordinates may coordinate their work autonomously ("horizontal coordination"). As explained in Section 2, the model could be easily adapted to cover this case.

[4] Balanceness of the hierarchy is implied, in my model, by the requirement that tasks are divided evenly among the members of the same workgroup. This assumption is discussed further in Section 2, where the model is presented.

agers. However, reports summarizing the relevant information must now be transmitted up the hierarchy to the top management, as in Radner (1993). Thus, upper-level managers will spend time both processing information (i.e., reading reports) and coordinating. This scenario is studied in Section 6.

I stress that the crucial distinction between production and information processing hierarchies revolves around whether or not upward reporting takes place within the organization. In particular, it may well be possible that the task involves some processing of information even in production hierarchies, so long as this information is not communicated to the senior management. Thus, for consistency and expositional ease, I will usually refer to the task as the processing of information, both in the production and the information processing scenario.

This paper studies how coordination responsibilities should be optimally allocated across layers taking into account the simple but important fact that communicating is inherently costly. The time managers spend on coordinating (which I term coordination costs), in fact, has to be computed as part of the organization's total working time and taken into account at the moment of making decisions. The organization's objective is to minimize a weighted average of the total wage bill and the cost of delay. The optimal level of communication results from a basic tradeoff between the savings on processing time due to less duplication and coordination costs. I stress that, as in the team-theoretic literature, no conflict of interest is assumed between the organization and its members, and managers will therefore simply maximize the organization's objective.

Although my analysis neglects the crucial role of incentives in organizations, it emphasizes a number of other important issues in the design of organizations, most notably the role of delay. To build some intuition, I first consider a simplified framework in which the span of control is constant across layers (a uniform hierarchy). In that setup, I show that in the optimum higher-level managers spend more time on communicating and coordinating than lower-level managers do. In production hierarchies, this is simply due to the fact that the effort of a senior manager influences a much greater portion of the hierarchy than the effort of a lower-level manager. Coordination is therefore more cheaply provided by higher-level managers. However, when managers must also transmit their information to their superiors (as in Section 6), a second asymmetry arises. In fact, provided that duplications made by subordinates can be detected and disregarded at no cost by their superiors, a subordinate's coordination efforts do not reduce (in form of shorter reports) his superior's information processing workload. By contrast, duplications made at higher levels always trickle down the hierarchy through the delegation process. Again, coordination effort is most valuable when exerted by the senior management, but now the reason is also that the benefits of coordination are higher at the top of the organization.

Empirically, the prediction that the time a manager spends on coordination activities increases as one moves up the hierarchy is strongly supported by the evidence. Mahoney et al. (1965), for instance, find that the percentage of managers who plan and the percentage of managers who coordinate as their main

function increase as one passes from low to middle levels in the hierarchy, and then again from middle to high.[5] Similarly, several other studies (see the surveys by Sayles (1964) and Guetzkow (1981), and the references therein) support the view that the managerial job entails a continuing effort to coordinate activities within the organization, and that planning and coordinating receive the greatest emphasis within top management. Importantly, these findings cannot entirely be attributed to the well-known fact that managers work longer hours the higher they are in the hierarchy, since they typically show that it is the percentage of time spent on planning and coordinating to total hours worked that is increasing with rank. Rather, the evidence suggests that the longer hours spent on these activities might be one of the reasons why higher-level managers work longer hours compared to lower-level managers.

A second result of this paper is that a shift towards granting employees broader decision authority – on the way work is coordinated, in the present model – has to be expected when reducing delay becomes more important. More precisely, the model shows that the ratio of communication effort exerted by a subordinate to the coordination effort exerted by his superior tends to increase with urgency and in the limit (i.e., when only delay matters) communication efforts are equalized across (managerial) layers. In this precise sense, therefore, increased urgency tends to 'empower' lower-level managers, relative to their superiors. Interestingly, the result seems broadly consistent with the view that innovative firms have more decentralized hierarchical structures than less innovative companies, and that globalization, by increasing competition (and therefore the need for fast decision-making), has had a major impact in forcing firms to restructure (Marin and Verdier (2003)).

Section 4 studies the more general case of balanced hierarchies, where the choice of the span of control across layers is also endogenous. Smaller spans of control near the top of the organization are shown to be typically beneficial as they reduce the number of managers in the hierarchy (keeping the number of workers fixed) and therefore coordination costs, particularly near the top where communication requirements are higher. Sufficient conditions are provided for the optimal span of control to be decreasing and coordination effort increasing as one travels up the hierarchy, with equalities holding if wages are negligible relative to the marginal cost of delay, thus generalizing the main result of Keren and Levhari (1979). These conditions turn out to be quite intuitive and, loosely speaking, essentially require communication effort and the span of control not to be 'too complementary' at reducing duplications. Furthermore, as for the time devoted to coordination activities, there seems to be some empirical support for the fact that the span of control is smaller near the top of the organization (e.g., Starbuck (1971), Gabraith (1977)), and Keren and Levhari

[5] Managerial jobs which are usually referred to as coordination activities by the empirical literature on human resource management (HRM) include the exchanging of information with people in the organization in order to relate and adjust programs, advising other departments, arranging meetings, etc. Planning activities such as work scheduling and programming are also examples of "coordination" as defined in the present model. The HRM literature also stresses the importance (in terms of time) of the planning and coordinating functions in actual organizations (e.g., Mahoney et al. (1965)).

(1979) also argue that, in military organizations, where wage costs are plausibly secondary compared to the utility of planning time saved, spans of control tend to be more uniform.

I also wish to emphasize the methodological contribution that this paper makes to the literature. Previous work on loss of control, including Keren and Levhari (1979,1983,1989), Qian (1994) and Meagher (2003), has relied heavily on the use of continuous approximations in the study of hierarchies. Van Zandt (1995), however, has shown that these approximations can be inaccurate, especially when the number of layers in the hierarchy is concerned. It is therefore worthy of mention that this paper extends previous results without relying on approximations but using instead interchange arguments which are standard in dynamic programming (see, e.g., Ross (1983)).

The present model can also be easily adapted to study the optimal choice of the managers' communicative skills in hierarchies. I consider a variant of the basic model in which managers devote the same amount of time to coordination activities, but the effectiveness of their effort depends on their ability. Not surprisingly, managers with strong interpersonal and communicative skills will typically be located at the top echelons of the hierarchy. However, the model also predicts that, as urgency increases, the optimal skill level of all managers should increase, especially at the lower and middle levels of the hierarchy. Empirically, this result suggests that firms operating in turbulent environments (and for which delay is presumably very costly) should hire more skilled managers and provide more training opportunities to their employees, especially at the bottom of the hierarchy, than companies operating in traditional sectors.

Finally, I consider another variant of the basic model (with constant span of control across layers) incorporating both a concern for delay and gains from specialization in information processing. I first show that the span of control and the number of levels in the hierarchy are, in a specific technical sense, substitutes, as they both increase the ability of the organization to process information concurrently by expanding its size. The focus of the analysis is on how these two variables vary together in the optimum as urgency increases. I show that if, in a sense made precise in the paper, delegation is mainly driven by specialization, then the span of control increases and the hierarchy flattens as reducing delay becomes more important. In a specific example, this condition is more likely to be fulfilled when delegation involves large fixed delay costs, gains from specialization are substantial and coordination costs are small. Taken together, my results offer a first formalization of the widespread view among practitioners that empowerment and delayering are to a large extent driven by the need for fast decision-making and, I hope, may shed some light on recent evidence on the changing nature of corporate hierarchies as documented, for instance, by Rajan and Wulf (2004).

1.1 Related Literature

This paper is related to several strands of the literature on hierarchies and organizations. It builds on recent contributions to the theory of information processing networks (e.g., Radner (1993), Bolton and Dewatripont (1994)), but the focus of the analysis is different. The information processing literature in fact focuses on the optimal design of the network and assumes that information arrives to the organization already disaggregated in the form of distinct batches of information. By contrast, the emphasis here is on coordinating the division of labor among workers but, for tractability, attention is restricted to balanced hierarchies.

This paper most directly contributes to a large literature on loss of control in organization (e.g., Williamson (1967), Calvo and Wellisz (1978), Keren and Levhari (1979)). Relative to previous contributions, besides the original focus on duplications, the present work pays special attention to the issue of delay. Considerations of delay are undoubtedly key in shaping organizational decisions and indeed several recent trends in the design of organizations have been attributed to a need for faster decision making and execution. Yet, it is fair to say that, with a few exceptions,[6] little work has been done to understand the impact of increased urgency on organizational design. This paper therefore fills an important gap in the literature and, in so doing, provides new insights on the issues of empowerment and delayering.

Lastly, this paper is related to a recent and fast-growing literature on the role of coordination within the firm (e.g., Dessein and Santos (2003), Harris and Raviv (2002), Rotemberg (1999)). Within this literature, the contribution most closely related to the present one is Hart and Moore (1999), which shows that coordinators (i.e., individuals whose tasks cover a large subset of assets) should appear higher in the chain of command than specialists (i.e., those with a narrow remit). However, Hart and Moore analyze hierarchy in terms of authority (which means that if i is above j in the hierarchy, then necessarily i has authority over j), whereas my work takes an information processing approach. The two papers are therefore best seen as complements rather than substitutes.

The remainder of the paper is organized as follows. Section 2 introduces the basic notation and assumptions of the model. Section 3 focuses on uniform hierarchies. Section 4 endogenizes the choice of the span of control and proves the main results of the paper. Section 5 focuses on delayering, while Section 6 studies information processing hierarchies. Section 7 concludes. All the omitted proofs are gathered in the Appendix.

[6] Keren and Levhari (1979) is a case in point. Beggs (2001) also characterizes the optimal design of the hierarchy where a trade-off exists between cost and delay and highlights a subtle relationship between increased costs of delay and decentralization. He shows that increased urgency need not always result in more decentralization because reducing the number of levels tasks have to pass through is only one of the methods an organization can employ to reduce delay. Another method, which may be cheaper, is to hire more low-ability workers to reduce the time spent queueing for attention.

2 The Model

This section introduces the main innovation of the paper, the duplication function, and discusses the key assumptions of the paper. A simple two-layer hierarchy is studied to illustrate the basic tradeoffs of the model. The general setup is then presented.

2.1 Duplication of Effort

Duplication of effort is modeled as follows. Consider a manager who must carry out a task of size $N \in \mathbb{R}_{++}$. As in Becker and Murphy (1992), it is assumed that any task can be subdivided into infinitely many subtasks. Suppose that the manager delegates and subdivide his task evenly among n subordinates (called a workgroup). Then duplications can occur as a result of a lack of coordination during the division of labor. Specifically, I posit that the actual total amount of information processed by the workers is not N but $ND(n,c) \geq N$, where $n \in \mathbb{N}$ is the span of control of the manager and $c \in [0,1]$ denotes the total communication effort exerted by the manager. Thus, $D(n,c)$ measures the amount of duplications in the organization or, equivalently, the lack of coordination among subordinates. Note also that the manager cannot retain part of the task for himself. This assumption is standard in models of loss of control and could be defended on the ground that managers have different skills than production workers,[7] but of course it would be worthwhile to relax it in future research.

Formally, the duplication function is defined as follows[8]

Definition (DF) For all $c \in [0,1]$ and n, the duplication function $D : \mathbb{N} \times [0,1] \to \mathbb{R}$ satisfies the following assumptions: (i) $D(n,c)$ is twice continuously differentiable in c, (ii) $D(n,c) \geq 1$, (iii) $D_c(n,c) < 0$ for all $2 \leq n \leq \bar{n}$ and (iv) $D(n,c) = n$ for all $n > \bar{n}$.

The first three assumptions impose mild restrictions on $D(\cdot,\cdot)$. In particular, (ii) just state that we are dealing with duplications, while (iii) says that communication reduces duplication of effort. The fourth assumption is convenient because it ensures that the optimum span of control is finite even in the extreme case where the organization only cares about delay, but could be easily relaxed. For instance, it would suffice to assume that duplications become sufficiently large as group size grows beyond a certain level, an assumption which is in line with a large managerial literature postulating an upper bound to the number of subordinates that can be effectively supervised.[9] Besides that, for most of the paper it will not be necessary to specify exactly how D changes as n changes, although one might reasonably expect

[7] Throughout this paper, however, all agents are assumed to be identical.

[8] By convention, subscripts in functions of more than one variable will be used to denote partial derivatives.

[9] Urwick (1956), for instance, relates the optimality of limited spans of control to bounds to "the number of items that the human brain can keep within its grasp simultaneously" (p.41). Psychological research also shows that disruption of communication is more likely to occur as the size of the group increases (see, e.g., Guetzkow (1981)).

duplications to increase monotonically with group size for fixed communication effort. Finally, I emphasize that this paper takes a reduced-form approach in modeling duplications. In the conclusion, however, I will sketch two possible ways in which the duplication function can be micro-founded.

2.2 A Two-Layer Hierarchy

In this subsection, I focus on the simple case of a two-layer production hierarchy and use this setup to illustrate the basic tradeoffs of the model.

I start with the trivial observation that processing information takes time. In particular, I assume that ηw is the time spent by a manager processing an amount w of information (his workload), where $\eta > 0$ measures the time necessary to process one unit of information. Planning and coordinating also take time. Specifically, let $\tau(c)$ be the time that a manager must spend on coordinating to achieve a communication level c.[10] $\tau(c)$ is assumed to be twice continuously differentiable, with $\tau(0) = 0$ and $\tau'(\cdot) > 0$. The assumption that $\tau(\cdot)$ is increasing simply reflects the fact that, if coordination has to be improved, the time spent on communicating must increase. c and $\tau(c)$ can be interpreted in at least two ways (see subsection 4.1 for a third interpretation in terms of ability). One possibility is that c denotes *downward communication* and $\tau(c)$ measures the amount of time that a superior spends communicating with his subordinates. Alternatively, c might denote *lateral communication and* $\tau(c)$ might measure the time each subordinate spends communicating with the other members of his workgroup. In the following, I will adopt the former interpretation; however, all the forthcoming results carry over (up to some changes in notation) to the latter scenario.

In addition to the variable communication cost $\tau(c)$, delegation also involves a fixed time cost. Specifically, I assume that an amount of time $b > 0$ must elapse before a manager can communicate with his subordinates. For instance, b may represent the time passing between when an e-mail is sent by a superior and when it is read by his subordinates, or the time necessary to organize and inform all participants about a meeting. By contrast, in those settings $\tau(c)$ might measure the time the superior spends writing that e-mail or attending that meeting.

The organization wishes to minimize the total cost of processing all the information. Two sources of costs are considered, the wage bill and the cost of delay. The wage bill includes the wages paid to the employees for the time they actually spend processing information and coordinating. In particular, as in Bolton and Dewatripont (1994), employees are not compensated for the time in which they remain idle, thus implicitly assuming that agents or the organization can use this idle time for other purposes. Delay is defined as the time the organization takes to process all the information, taking into account that some

[10] The time and effort it takes to the subordinates to absorb these instructions is assumed to be zero. This is similar to the convention adopted in the information processing literature that all the costs of reporting are borne by superiors in the form of reading time.

tasks might be processed in parallel. Thus, for instance, in a one-layer hierarchy, the wage bill is $\omega\eta N$ (where ω denotes the wage for unit of time and with no loss of generality is normalized to unity in the sequel) and delay is given by ηN. The total cost of processing N units of information using only one worker is given by $\eta N + \lambda\eta N$, where $\lambda > 0$ denotes the weight attached to (the cost of) delay relative to wage costs (or, alternatively, the constant marginal cost of delay). I emphasize that the assumption that the cost of delay is linear in delay is just for simplicity. Indeed, once the general model has been described, I will explain how many of the special features of this model can be relaxed without affecting the results of the paper.

The next step is to consider a two-layer hierarchy. In this scenario, the top manager (at layer 1) delegates the task to a number of subordinates (say n) at layer 2 who then process all the information. Because of duplications, the size of the task at the second level is $ND(n, c)$, and $\tau(c)$ denotes the time the top manager spends on coordinating. Provided tasks are divided evenly among subordinates, a subordinate's workload is $w = ND(n, c)/n$ and $\eta ND(n, c)/n$ is the time he spends processing information. The total cost of a two-layer hierarchy (information processing costs plus delay) is therefore given by

$$T(n, c, 2) = [\eta ND(n, c) + \tau(c)] + \lambda [\eta ND(n, c)/n + \tau(c) + b]. \tag{1}$$

Note that delay includes only the time spent processing information by *one* manager at layer 2, since the managers at the same level in the hierarchy work concurrently (this is a standard assumption in the literature). Furthermore, the time elapsing before a manager can communicate with his subordinates, b, only affects delay, since by assumption the organization does not pay its employees for the time they remain idle.

The problem of the two-layer hierarchy is to minimize (1) with respect to c and n. Before proceeding further, I illustrate the key tradeoffs of the model in this simple setup. Assume for simplicity interior solutions. The first order condition with respect to c yields[11]

$$(1 + \lambda)\tau'(c^*) = -\eta N(1 + \lambda/n)D_c(n, c^*). \tag{2}$$

Clearly the tradeoff here is between higher coordination costs and more duplication of effort. Furthermore, assuming just for the moment that the span of control can be treated as a continuous variable, the first order condition with respect to n yields

$$-\lambda \left.\frac{\partial(D(n, c)/n)}{\partial n}\right|_{c=c^*, n=n^*} = D_n(n^*, c^*). \tag{3}$$

Thus, if $\frac{\partial(D(n,c)/n)}{\partial n} < 0$, the benefit of a larger span of control is that bottom-level managers have smaller workloads and therefore delay is reduced. The drawback is that duplications and hence wage costs may increase ($D_n(n, c) > 0$). The latter effect can be seen as a formalization of Simon and al.'s insight that

[11] Throughout the paper, stars will denote optimal values.

"the maximum size of an effective unit is limited basically by the ability of that unit to solve its problems of internal communication" (1950, p.131). Indeed, $D_n(n,c) > 0$ seems a particularly natural assumption since, for given c, duplications will tend to rise with group size as the top manager spends on average less time communicating with each subordinate. Furthermore, even for fixed $\frac{c}{n}$, each unit of communication may become less effective as group size grows because misunderstandings and other coordination failures are all more likely to occur in large groups (Guetzkow (1981)).

Finally, the (optimal) two-layer hierarchy is preferred to a one-layer/one-manager hierarchy whenever the total cost of the former (wage bill and delay) is lower than the total cost of the latter:

$$\lambda N\eta \left[1 - \frac{D(n^*,c^*)}{n^*}\right] > N\eta \left[D(n^*,c^*) - 1\right] + (1+\lambda)\tau(c^*) + \lambda b. \tag{4}$$

That is, it must be the case that the reduction in delay due to parallel information processing, which is proportional to $1 - \frac{D(n^*,c^*)}{n^*}$, outweighs duplication and coordination costs. Equation (4) is more likely to be fulfilled when duplications and coordination costs are small and λ,η and N are large.

2.3 The Problem of the Hierarchy

The two-layer model studied above can be generalized to hierarchies with an arbitrary number of layers. Consider an organization composed of $L \geq 2$ layers. Tasks are delegated as follows. The top manager, at level 1, delegates his tasks (of size N) to n_1 subordinates and exert communication effort c_1. Because of duplications, layer 2 total task (that is, the information managers at layer 2 believe they are responsible for) is $ND(n_1,c_1)$. I assume that delegated tasks are divided evenly among the members of the same workgroup, which implies that the size of an individual task at layer 2 is $ND(n_1,c_1)/n_1$. Each manager at layer 2 then delegates his task to n_2 managers at layer 3. Thus, a workgroup's task at layer 3 is $ND(n_1,c_1)D(n_2,c_2)/n_1$, and layer 3 total and individual tasks are, respectively, $ND_1(n_1,c_1)D(n_2,c_2)$ and $\frac{ND_1(n_1,c_1)D(n_2,c_2)}{n_1 n_2}$. In general, layer l total task is $N\prod_{i=1}^{l-1} D(n_i,c_i)$ and layer l individual tasks are $N\prod_{i=1}^{l-1}\frac{D(n_i,c_i)}{n_i}$, $l = 2,..,L$. At layer L, tasks cannot be delegated further and the information is processed: a worker's workload is therefore $w_L = N\prod_{i=1}^{L-1}\frac{D(n_i,c_i)}{n_i}$.[12] Importantly, note that this expression only depends on communication efforts and spans of control through the multiplicatively separable function $\prod_{i=1}^{L-1}\frac{D(n_i,c_i)}{n_i}$.

Some features of this setup are worthy of mention. The most important remark concerns the assumption that tasks are divided evenly among subordinates. This assumption greatly simplifies the model as it allows me to focus on balanced hierarchies. In fact, if all managers at the same layer face the same problem/task, one can assume with (essentially) no loss of generality that they will also make the same decisions, and in

[12] An important distinction in this model is between tasks and workloads. Workloads refer to the size of the task which is actually processed. Tasks, by contrast, refer to a manager's assignment, which could be delegated and therefore not necessarily processed. Of course, at the bottom level, tasks and workloads coincide.

particular that they will choose the same communication efforts and spans of control. Balanced hierarchies are especially attractive because they resemble, at least to a first approximation, actual managerial charts; furthermore, their simplicity allows me to concentrate on the main theme of the paper, the coordinating function of the management. Nevertheless, in future it would certainly be desirable to fully endogenize the division of labor within the organization, for instance by assigning to each agent a task of a potentially different measure.[13]

A second feature of the model that must be emphasized is that the duplication function is the same at every level of the hierarchy. This is consistent with the assumption that the organization's task is perfectly divisible and homogeneous. However, in reality the job of the senior management is often qualitatively different from the job of lower-level managers. In my final remarks, I will briefly mention how asymmetries in the difficulty/complexity of the task across layers may affect the conclusions of the paper.

Third, I stress that although a superior only coordinates the division of labor within the workgroup he supervises, coordination among workgroups is guaranteed by the efforts of the upper-level managers. Thus, for instance, even if there was no duplication within workgroups at the bottom level of the hierarchy, there may still be duplications among workgroups as a result of poor coordination at previous layers.

Returning to the model, note that the number of managers at layer l is $\prod_{i=0}^{l-1} n_i$ (by convention $n_0 = 1$ refers to the top manager). Total processing time is therefore $\eta \prod_{i=0}^{L-1} n_i w_L = \eta N \prod_{i=1}^{L-1} D(n_i, c_i)$. Total working time (processing time and coordination costs) and the wage bill are given by

$$WB = \eta N \prod_{i=1}^{L-1} D(n_i, c_i) + \sum_{l=1}^{L-1} \prod_{i=0}^{l-1} n_i \tau(c_l). \tag{5}$$

However, organizations cannot simply be evaluated on the grounds of how cheaply they process information, but must also be able to swiftly react to new information. The notion of (the cost of) delay introduced in the previous paragraph can readily be generalized to multilayer hierarchies as follows:

$$C(Delay) = \eta N \prod_{i=1}^{L-1} \frac{D(n_i, c_i)}{n_i} + \sum_{l=1}^{L-1} \tau(c_l) + (L-1)b. \tag{6}$$

As equation (6) makes clear, delay is the time it takes the hierarchy to process all information. In particular, this expression considers only one employee at each level of the hierarchy since managers at the same layer work in parallel.

The problem of the hierarchy is to minimize a weighted sum of the wage bill and the cost of delay with respect to the number of layers , the span of control and the communication efforts at each level of the hierarchy. Let $\mathbf{c} = (c_1, ..., c_{L-1})$ denote the vector of communication efforts exerted by the managers at

[13] In a different setup, Radner (1993) has shown that the unbalanced hierarchies he termed 'reduced trees' are optimal. Interestingly, Radner (1993) and Van Zandt (2004) also provide conditions under which optimal hierarchies are 'approximately balanced'.

the various layer of the hierarchy. Similarly, let $\mathbf{n} = (n_1, ..., n_{L-1})$, where n_l is the span of control at layer l. The problem of the hierarchy is

$$\min_{\mathbf{n},\mathbf{c},L} T(\mathbf{n}, \mathbf{c}, L) = \eta N \prod_{i=1}^{L-1} D(n_i, c_i) + \sum_{l=1}^{L-1} \prod_{i=0}^{l-1} n_i \tau(c_l) + \lambda \left(\eta N \prod_{i=1}^{L-1} \frac{D(n_i, c_i)}{n_i} + \sum_{l=1}^{L-1} \tau(c_l) + (L-1)b \right) \quad (7)$$

for $L \geq 2$. As before, $\lambda \in \mathbb{R}_{++}$ denotes the weight attached to delay (or, alternatively, the marginal cost of delay) relative to the wage bill. When only delay matters, I write $\lambda = \infty$ to say that the weight attached to the wage bill, ω, is zero. Of course, in that case the problem of the hierarchy reduces to minimizing (6) with respect to $\mathbf{c}$, $\mathbf{n}$ and L.

Proposition 1 *A solution to programs (7) and (6) exists. Moreover, in the optimum* $\bar{n} \geq n_l^* \geq 2$ *for all* $l = 1, ..., L-1$.

The result that in the optimum every manager must have at least two subordinates follows from the fact that if a task is delegated to a single subordinate, no parallel processing and hence no reduction in delay would occur. Thus it would not be optimal to delegate in the first place. Furthermore, no manager would want to supervise too many subordinates ($n > \bar{n}$), since by assumption that would impair the manager's ability to do his job and cause a lot of duplications (of course $\bar{n}$ could be 'large'). As argued above, this seems quite realistic because in practice only a limited number of subordinates can be effectively supervised by a single manager. Finally, I wish to emphasize that several restrictive assumptions implicit in (7) can easily be relaxed without affecting the conclusions of the paper. In particular, all the main results carry over if

- The cost of delay is any strictly increasing function of delay, not necessarily linear.[14]
- The time necessary to process one unit of information is any increasing function of a worker's workload, w_L, so that individual processing time is given by $\eta(w_L)w_L$, with $\eta'(\cdot) \geq 0$.[15] Thus, the model can easily accommodate gains from specialization in information processing as modeled in Becker and Murphy (1992), or overload costs.
- Managers and production workers receive different hourly wages, ω_m and ω_p, instead of a common salary ω per unit of time. In that case, the condition $\lambda = \infty$ should be read as saying that both ω_m and ω_p are negligible compared to the weight attached to delay.

[14] This is true except for the results in Section 5. Even in that section, however, the results do not crucially rely on the linearity of $C(\cdot)$, and less restrictive conditions could be given.

[15] Again, the exception is Section 5 where I will also require $\eta(\cdot)$ to be convex, so that returns from specialization are decreasing.

3 Optimal Communication in Uniform Hierarchies

To build some intuition for the results, this section focuses on the simpler case in which the span of control is constant across layers and equal to n (a uniform hierarchy). The analysis of the general case (7) is postponed to the next section.

For any given $L \geq 3$ and $\bar{n} \geq n \geq 2$, the problem of the (uniform) hierarchy is

$$\min_{\mathbf{c}} T(\mathbf{c}) = \eta N \prod_{i=1}^{L-1} D(n, c_i) + \sum_{l=1}^{L-1} n^{l-1} \tau(c_l) + \lambda \left(\eta N \prod_{i=1}^{L-1} \frac{D(n, c_i)}{n^{L-1}} + \sum_{l=1}^{L-1} \tau(c_l) + (L-1)b \right). \tag{8}$$

The following two propositions characterize the optimal assignment of coordination responsibilities in uniform hierarchies.

Proposition 2 *Let $\lambda \in \mathbb{R}_{++}$ and $\mathbf{c}^* = (c_1^*, c_2^*, ..., c_{L-1}^*)$ be a vector of communication efforts that solves program (8). Then $c_1^* \geq c_2^* \geq ... \geq c_{L-1}^*$.*

Proof. Assume instead that $\tilde{\mathbf{c}} = (\tilde{c}_1, \tilde{c}_2, ..., \tilde{c}_{L-1})$, with $\tilde{c}_i < \tilde{c}_j$ for some $i < j$, solves (8) Then it is possible to improve upon $\tilde{\mathbf{c}}$. Indeed, let $\check{\mathbf{c}}$ be a vector such that $\check{c}_l = \tilde{c}_l$ for all $l \neq i, j$ and $\check{c}_i = \tilde{c}_j$ and $\check{c}_j = \tilde{c}_i$. Note that $T(\tilde{\mathbf{c}}) - T(\check{\mathbf{c}}) = n^{i-1}\tau(\tilde{c}_i) + n^{j-1}\tau(\tilde{c}_j) - n^{i-1}\tau(\tilde{c}_j) - n^{j-1}\tau(\tilde{c}_i) = (n^{j-1} - n^{i-1})\left[\tau(\tilde{c}_j) - \tau(\tilde{c}_i)\right] > 0$. Thus $\tilde{\mathbf{c}}$ cannot be optimal. This yields a contradiction. ■

Thus, in the optimum senior managers work and are hence paid more than junior managers. Loss of control and lack of coordination also become relatively more severe as one goes down the hierarchy since $D(n, c_{l-1}^*) \leq D(n, c_l^*)$ for all n and l. Finally, I stress that the interchange argument used in this proof (and in many subsequent proofs) crucially relies on the fact that a worker's workload $w_L = N \prod_{i=1}^{L-1} \frac{D(n_i, c_i)}{n_i}$ only depends on c and n in a multiplicatively separable fashion. Assuming more general cost of delay functions or that η depends on w_L would clearly not affect the results.

Proposition 3 *If $\lambda = \infty$ and $\mathbf{c}^*$ minimizes (6), (i) so does any permutation of $\mathbf{c}^*$. In particular, $c_1^* = c_2^* = ... = c_{L-1}^*$ whenever the solution is unique. (ii) Moreover, if $-\frac{\tau'(c)}{D_c(n,c)}$ is increasing in c for all n,*[16] *then $c_1^* = c_2^* = ... = c_{L-1}^*$.*

Proof. (i) When only delay matters, the problem of the hierarchy is

$$\min_{\mathbf{c}} \eta N \prod_{i=1}^{L-1} \frac{D(n, c_i)}{n^{L-1}} + \sum_{l=1}^{L-1} \tau(c_l) + (L-1)b. \tag{9}$$

Claim (i) is then obvious since all coordination efforts enter this expression symmetrically.

[16] Throughout this paper when I say that something 'is increasing', 'increases' or 'rises', I mean to say that it is nondecreasing. Most of the inequalities in this paper are weak, so this convention leads to less awkwardness. When I want an inequality to be strict, I will say so explicitly, as in 'strictly increasing', etc.

(ii) Assume by contradiction that $\tilde{\mathbf{c}} = (\tilde{c}_1, \tilde{c}_2, ..., \tilde{c}_{L-1})$ solves (9), with $\tilde{c}_i > \tilde{c}_j$ for some $i < j$ (the reasoning for $\tilde{c}_i < \tilde{c}_j$ is similar). Let $\mathbf{x} = (x_1, ..., x_{L-1})$ with $x_i = \frac{D(n,c_i)}{n}$. Then (9) can be rewritten as $\min_{x_1,...,x_{L-1}} \left\{ N\eta \prod_{l=1}^{L-1} x_l + \sum_{l=1}^{L-1} f(x_l) + (L-1)b \right\}$ where $c_l = D^{-1}(n, nx_l)$ (by assumption D is strictly decreasing in c on the 'relevant domain' $2 \leq n \leq \bar{n}$) and $f(x_l) = \tau[D^{-1}(n, nx_l)]$. Denote $\tilde{x}_i = \frac{D(n,\tilde{c}_i)}{n}$. Now consider a vector $\breve{\mathbf{x}}$ whose elements are the same as in $\tilde{\mathbf{x}}$ except for the fact that $\breve{x}_i = k\tilde{x}_i$ and $\breve{x}_j = \frac{1}{k}\tilde{x}_j$. Intuitively, starting at $k = 1$, as k increases, the difference between $\breve{x}_j$ and $\breve{x}_i$ gets smaller since by assumption $\tilde{x}_i < \tilde{x}_j$. Define $g(k) = N\eta \prod_{l=1}^{L-1} \tilde{x}_l + \sum_{l\neq i,j}^{L-1} f(\tilde{x}_l) + f(k\tilde{x}_i) + f(\frac{1}{k}\tilde{x}_j) + (L-1)b$ to be the value of (9) at $\breve{\mathbf{x}}$ as a function of k. Clearly, if $g'(1) < 0$, then it is possible to improve upon $\tilde{\mathbf{c}}$ and therefore $\tilde{c}_i > \tilde{c}_j$ cannot be optimal. Note that $g'(1) = f'(\tilde{x}_i)\tilde{x}_i - f'(\tilde{x}_j)\tilde{x}_j < 0$ provided $f'(\tilde{x}_i) \leq f'(\tilde{x}_j)$ (since $\tilde{x}_i < \tilde{x}_j$), that is, if $f'(x)$ is increasing in x. By the Inverse Function Theorem, we get $f'(x) = \frac{\tau'[D^{-1}(n,nx)]}{D_c(n,D^{-1}(n,nx))} = \frac{\tau'(c)}{D_c(n,c)}$,[17] which by assumption is decreasing in c. Claim (ii) then immediately follows from the fact that $x = \frac{D(n,c)}{n}$ is decreasing in c. ■

The assumption that, for all n, $-\frac{\tau'(c)}{D_c(n,c)}$ is increasing in c is satisfied, for instance, whenever $\tau''(\cdot) > 0$ and $D_{cc}(\cdot,\cdot) = 0$ or $\tau''(\cdot) = 0$ and $D_{cc}(\cdot,\cdot) > 0$. It postulates decreasing (net) returns to coordination and implies that coordination is more valuable when it is scarce. Indeed, in practice people typically deal with the most important issues immediately, and tie up details at a later stage. Functional forms consistent with all the assumptions on $\tau(c)$ and $D(n,c)$ are, for instance, either $\tau(c) = \frac{c}{1-c}$ or $\tau(c) = \alpha c^\beta$, $\alpha > 0$, $\beta > 1$, and $D(n,c) = n^{1-c}$.

Example 1 Let $D(n,c) = n^{1-c}$ and either $\tau(c) = \frac{c}{1-c}$ or $\tau(c) = \alpha c^\beta$, $\alpha > 0$, $\beta > 1$, and assume interior solutions. Then program (8) is strictly convex and the unique optimal solution satisfies $c_1^* > c_2^* > ... > c_{L-1}^*$. Furthermore, $\partial\left(\frac{c_l^*}{c_{l+1}^*}\right)/\partial\lambda < 0$ and $\lim_{\lambda\to\infty} \frac{c_l^*}{c_{l+1}^*} = 1$ for all $l \leq L-2$.

The intuition for Proposition 2 is simple: In the optimum higher-level managers coordinate more than lower-level managers since the former exert their influence on a much greater portion of the hierarchy than the latter do. Consider the wage bill in isolation. Although the various duplication functions (and hence the coordination efforts) enter into the wage bill symmetrically, the individual (per capita) contribution of managers at different layers is different. Indeed, there are fewer and fewer managers as one goes up the hierarchy, and that difference grows exponentially. Thus the actions of higher-level managers have a much greater impact on the performance of the organization than those taken by their subordinates.[18] By contrast, the cost of delay treats individual working time at each layer symmetrically since only the contribution of one manager at each layer has to be considered. The second part of Proposition 3, in

[17] $\tau(c)$ is in fact strictly increasing by assumption and $D^{-1}(n,nx)$ is strictly decreasing in x on $2 \leq n \leq \bar{n}$.

[18] This is a purely cost-driven effect: coordination is just more cheaply provided near the top of the organization. In information processing hierarchies, however, it will be shown that not only are the costs of coordination lower near the top, but also the benefits may be larger.

particular, shows that when only delay matters, coordination efforts are equalized across layers. Since the optimal level of coordination results from a combination of these two effects, the introduction of delay tends to weaken the wage-cost effect but cannot cancel it, except for the case where only delay matters.

Clearly, Proposition 2 provides a rationale for the empirical regularity that the time a manager spends in planning and coordination activities generally increases with rank in the hierarchy, especially when the difference in level between managers is high. Proposition 3 and Example 1, instead, may be useful to understand the recent trend towards empowerment. According to the model, in fact, a shift towards granting employees broader decision authority – on the way work is organized, in the present case – has to be expected when the weight attached to the cost of delay increases relative to the weight attached to the wage bill. Together, these results can be interpreted as formalizing the presence and indeed the optimality of overloads at the top levels of the hierarchy, overloads which however are suboptimal from a purely delay minimizing point of view.

The reorganization of Pepsi in 1988 provides a nice illustration of how demands for faster information processing driven by changes in the external business environment can lead to greater decentralization of the decision making process (Besanko et al., 2000). For Pepsi, a key change was the emergence of large regional supermarket chains. These in fact often operated in large territories, whereas Pepsi's existing structures gave nobody region-wide authority over pricing. The problem was that regional executives, when faced with requests for promotions or special pricing deals, often disagreed over the appropriate strategy to follow. Their disputes had then to be resolved at the top levels of the hierarchy, which were forced to become involved in region-level pricing and promotion decisions. Not surprisingly, this impaired Pepsi's ability to respond and put it at a competitive disadvantage in a market that demands nimble marketing responses to fast-changing circumstances. The solution was found by reorganizing the hierarchical chain of command in a new, geographically oriented matrix structure. This structure created the position of area general manager, who had final authority for operational decisions, such as pricing and promotions, within areas that were roughly the size of the territories of the large supermarket chains. Thus, the task of coordinating previously semiautonomous geographic areas was delegated down the hierarchy to an appositely created new managerial figure in response to an increase in the cost of delay driven by a change in the organization's business environment.[19]

The following proposition captures the intuition that, in large hierarchical organizations, the top management is forced to spend a lot of time in coordination activities. Denote by $c_1^*(L)$ the optimum value of

[19] Boeing is another case in point. According to J.R. Galbraith (1977), "After 1964 the problem facing Boeing was not to establish a market but to meet the opportunities remaining as quickly as possible. [...] Now a delay of a few months would result in canceled orders and fewer sales" (p.191). As a result, instead of referring a problem upward in the hierarchy, coordination responsibilities were delegated to task forces and liaison groups of designers and engineers, who had to solve the problem at their own level, contacting and cooperating with peers in those departments affected by the new information.

c_1 in a L-layer hierarchy.

Proposition 4 *Consider program (8).* $c_1^*(L) \to 1$ *as* $L \to \infty$.

Two remarks are in order. First, the fact that $\lambda < \infty$ (which is implicit in (8)) is important. When only delay matters, in fact, it may well be the case that $c_1^*(L)$ decreases in L. Second, and more importantly, note that when $\tau(1) = \infty$ (which is true for instance if $\tau(c) = \frac{c}{1-c}$), then increasing indefinitely the number of levels in the hierarchy requires the top manager to bear arbitrarily large coordination costs. This model is therefore consistent with the view that the coordinating role of the top management is the main limiting factor of the size of the firm, and with Herbert Simon's claim that "attention is the chief bottleneck in organizational activities, and the bottleneck becomes narrower and narrower as we move to the tops of organizations, where parallel processing capacity becomes less easy to provide without damaging the coordinating function that is the prime responsibility of these levels" (Simon, 1976, p.294). Interestingly, the result holds despite the fact that coordination efforts are strategic substitutes (indeed, one can easily check that, by coordinating more, lower-level managers reduce the need for coordination at higher levels). Thus, Proposition 4 qualifies and extends the view that coordination limits the size of the firm because some tasks, most notably the coordinating function of the executive, cannot be delegated (see, e.g., Gifford (1992)).

4 Main Results: Communication vs. Span of Control

In this section I consider the general case in which the span of control can differ across layers and is chosen to minimize the organization's total costs. Starbuck, in particular, writes that "the span of control was supposed to be smaller near the top of the management hierarchy than near the bottom, because there was greater need for coordination near the top" (1971, p.88). This section studies how communication effort and the span of control interact to reduce duplications, and establishes conditions under which Starbuck's claim holds.

Let the problem of the hierarchy be given by (7). This problem is considerably more complex than the previous one since now it involves $2(L-1)+1$ endogenous variables, L of which can only take integer values. Nevertheless, a partial characterization of the optimal solution can still be obtained.

Proposition 5 *For all* $L \geq 3$, *let* $(\mathbf{n}^*, \mathbf{c}^*) = ((n_1^*, c_1^*), ..., (n_{L-1}^*, c_{L-1}^*))$ *be a solution to (7). There are no* $k, h \geq 1$ *such that* $c_k^* < c_{k+h}^*$ *and* $n_k^* > n_{k+h}^*$, $k + h \leq L - 1$. *Moreover, if* $n_k^* = n_{k+h}^*$, *then* $c_k^* \geq c_{k+h}^*$. *If* $c_k^* = c_{k+h}^*$, *then* $n_k^* \leq n_{k+h}^*$.

Proposition 5 generalizes Proposition 2, for now it suffices to know that $n_k^* \geq n_{k+h}^*$ to conclude that $c_k^* \geq c_{k+h}^*$. More importantly, Proposition 5 provides some support for the view that coordination in

organizations is more cheaply provided when communication efforts are higher and spans of control are smaller at the top of the hierarchy. Indeed, the proof exploits the fact that, so long as $c_k < c_{k+h}$ and $n_k \geq n_{k+h}$ (or $c_k \leq c_{k+h}$ and $n_k > n_{k+h}$), one can always reduce the total wage cost of coordination $\sum_{l=1}^{L-1} \prod_{i=0}^{l-1} n_i \tau(c_l)$ while keeping both delay and the workers' information processing workload constant by simply interchanging (n_k, c_k) with (n_{k+h}, c_{k+h}). However, the result falls short of showing that both small spans of control and more communication effort are always optimal near the top of the organization. Indeed, as the following proposition shows, to do so it is important to specify how communication effort and smaller spans of control interact to reduce duplications.

Proposition 6 *If $D(n,c)$ is log-supermodular,*[20] *then $c_k^* \geq c_{k+h}^*$ and $n_k^* \leq n_{k+h}^*$.*

The importance of this result lies in the fact that it highlights the role of complementarity between communication and span of control; in particular it requires c and n not to be 'too complementary' at reducing duplications. This is quite intuitive: If the incentive to exert effort increased dramatically with the number of subordinates, the result would never hold. Nevertheless, the requirement that $D(n,c)$ be log-supermodular is very restrictive, as it rules out the possibility that communication becomes (strictly) more valuable as the size of a workgroup grows. Importantly, however, log-supermodularity is only a sufficient condition, which becomes less and less stringent as the hierarchical gap h between two managers grows. To see that, note that the wage cost of one additional hour of communication at layer l grows exponentially as one goes down the hierarchy since $\prod_{i=0}^{l-1} n_i \geq 2^{l-1}$. Thus, provided that the organization is concerned about wage costs and h is large (for fixed N and L), it is realistic to expect $c_k^* \geq c_{k+h}^*$, regardless of whether n_k^* is greater or smaller than n_{k+h}^*.[21] A simple example may also help clarify the magnitude of this effect. Assume for instance that the managers at layer 1 and 6 of an organization spend the same time on communicating, and let the span of control of the managers at layer 1 be six, and the span of control of all the other managers be eight. Then the cost of a marginal increase in communication effort is almost 200,000 times smaller at layer 1 than it is at layer 6, since there are far more managers at the lower level. It is therefore likely that $c_1^* \geq c_6^*$, even if communication may be more 'effective' at the lower level, where the span of control is larger.

Next, consider the problem of the hierarchy when only delay matters ($\lambda = \infty$).

Proposition 7 *Suppose (6) has a unique solution,*[22] *and denote it by $(\mathbf{n}^*, \mathbf{c}^*)$. Then $c_1^* = c_2^* = ... = c_{L-1}^*$ and $n_1^* = n_2^* = ... = n_{L-1}^*$.*

[20] Log-supermodularity of a positive function $h(x,t)$ implies that the relative returns, $h(x_H,t)/h(x_L,t)$, are increasing in t for all $x_H > x_L$. (For a formal definition, see Athey (2002).) A simple example of a log-supermodular function is $D(n,c) = v(n)d(c)$. $D(n,c) = n^{1-c}$ is log-submodular.

[21] A similar reasoning applies for the span of control.

[22] Due to integer constraints, it is difficult to solve program (6) and to show that a solution is unique. However, ignoring such constraints, one can assume that (6) is strictly convex. The true solution will then be one of the integer solutions nearest to the unique solution found and if the organization is large the relative error will be small. Moreover, if $D(n,c)/n$

Proposition 6 and 7 show that Keren and Levhari's (1979) main result, that the span of control is decreasing as one travels up the hierarchy, with equality holding if wages are negligible relative to the marginal cost of delay, can be derived as a special case of this model. More importantly, these results seem in line with the empirical evidence. I have already mentioned a few studies showing that planning and coordinating receive the greatest emphasis within top management. There also seems to be some empirical support for the idea that the span of control should be smaller near the top of the organization (e.g., Starbuck (1971), Gabraith (1977)). Table 3.6 in Galbraith (1977, p.48), in particular, shows that in the production departments of United States and Canadian oil refineries, the span decreases as one passes from second level (Foreman) to the third (General Foreman), and then again from the third to the fourth (Superintendent). Interestingly, Keren and Levhari (1979) also argue that, in military organizations, where the cost of staffing the hierarchy is secondary compared to the utility of planning time saved, spans of control tend to be more uniform. However, it must be acknowledged that these shreds of evidence are far from conclusive and different models, in particular Beggs (2001) and Calvo and Wellisz (1979), have found the opposite result, namely that top managers control more direct subordinates than junior managers do.

4.1 Communicative Skills

So far this paper has focused on situations in which all managers are equally talented. I showed that higher-level managers will typically spend more time on coordination activities than lower-level managers simply due to their different positions in the hierarchy, not because of inherent differences in ability. Indeed, since all our managers are identical, it does not really matter which hierarchical position a manager is allocated to. However, when differences in ability are taken into account, this indifference result is no longer likely to be valid, and several models have in fact shown that the more able agents should be assigned to the higher levels of the hierarchy. These models by and large focus on talent as either information processing speed (e.g. Prat (1997)) or as the ability to monitor subordinates (e.g. Calvo and Wellisz (1979)). Casual empiricism, on the other hand, suggests that the ability to coordinate effectively is very important at the top levels of the hierarchy, probably more important than the ability to process information rapidly. Indeed, it is often required that successful candidates for executive positions have 'strong interpersonal and communicative skills, a collegial management style', 'ability to organize work and demonstrated ability to work in harmony with people', 'ability to communicate effectively'.[23] Within the economics literature, Williamson emphasizes the same point when he writes:

is decreasing in n over the relevant range, then $n_1^* = ... = n_{L-1}^* = \bar{n}$.

[23] Furthermore, the Bureau of the Census (1995) shows that, when recruiting new production staff, US employees rank communication skills above, e.g., previous work experience, recommendations from current employees, years of schooling and grades. According to the employees, the only factor in making hiring decisions which was more important than communication skills was "applicant's attitude".

> The bounds of rationality here take the form of language rather than computational limits and evidently vary among individuals. If the specialization of labor is feasible, those whose rationality limits are less severely constrained than others are natural candidates to assume technical, administrative or political leadership positions – which is to say that hierarchy can emerge on this account (1975, p.24).

This subsection presents a simple variant of the problem of the hierarchy which endogenizes the choice of a manager's communicative skills, or ability. To keep things simple, assume that every manager spends the same amount of time a coordinating the work of his direct subordinates (a may be interpreted as the length of a normal working day).[24] However, managers differ in their communicative skills. Specifically, let $c \in [0,1]$ index communicative skills and $\tau(c)$ be the wage paid to a manager of ability c working a hours. Clearly τ is increasing in c since higher-ability managers must be paid more than their lower-ability counterparts. Furthermore, let $D(n,c)$ measures the amount of duplications that arise during the delegation process when a superior's ability is c and n is the size of the workgroup. Obviously, the abler the superior, the fewer the duplications. The modified problem of the hierarchy is

$$\min_{\mathbf{n},\mathbf{c},L} T(\mathbf{n},\mathbf{c},L) = \eta N \prod_{i=1}^{L-1} D(n_i,c_i) + \sum_{l=1}^{L-1}\prod_{i=0}^{l-1} n_i \tau(c_l) + \lambda\left(\eta N \prod_{i=1}^{L-1}\frac{D(n_i,c_i)}{n_i} + (L-1)(a+b)\right). \qquad (10)$$

Note that the only difference between program (10) and (7) is that the time cost of coordinating subordinates at any given layer l is now $a+b$ and not $\tau(c_l)+b$. Notwithstanding this, it is not hard to see that the results of the previous sections carry over to the new framework. In particular, and in line with Williamson's claim, managers with strong interpersonal and communicative skills will typically be found at the top echelons of the hierarchy. In addition, however, the fact that the time cost of coordinating subordinates is now independent of c yields a novel implication. When only delay matters, in fact, there is now no tradeoff between fewer duplications and higher wages, and therefore in the optimum $c_1^* = ... = c_{L-1}^* = 1$. Thus, empirically, the result suggests that, as urgency increases, the average quality of the management should increase, especially at the lower and middle levels of the hierarchy. In particular, the model makes the testable prediction that firms operating in turbulent environments (and for which delay is presumably very costly) should hire managers of higher ability and provide more training opportunities to existing employees, especially at the bottom of the hierarchy, than companies operating in mature sectors.

5 The Flattening Hierarchy

This section focuses on relationship between the optimal number of layers and the span of control. My goal is to investigate the widespread belief among practitioners that 'delayering' (i.e., the reduction in

[24] This is of course a strong assumption since typically managers work longer hours the higher they are in the hierarchy.

the number of formal levels in an organization) is a powerful instrument to speed up decision making. Jeffrey Immelt, the CEO of General Electric, for instance, motivated the decision to shorten the chain of command at GE as follows: "The reason for doing this is simple—I want more contact with the financial services teams....With this simplified structure, the leaders of these four businesses will interact directly with me, enabling faster decision making and execution."[25] Furthermore, Rajan and Wulf (2004) show that in recent years US corporations have experienced both a reduction in the number of formal layers and an increase in the number of managers directly reporting to the top management.

In this section, a very stylized model incorporating both a concern for delay and gains from specialization in information processing is presented. I show that if, in a sense made precise below, delegation is mainly driven by specialization (rather than delay), as urgency increases, the organization becomes flatter and the span of control increases. Using a specific information processing technology, I also show that this condition is more likely to be fulfilled the greater the fixed delay caused by delegation and the strength of the gains from specialization, and the smaller the coordination costs.

To focus on the span of control and the number of levels in the hierarchy, a number of simplifying assumptions are made. Communication effort is assumed to take only two possible values, $c \in \{0,1\}$, with $\tau(1) = \bar{\tau} > 0$ and $\tau(0) = 0$. When managers exert effort ($c = 1$), coordination is perfect and hence there are no duplications: $D(n,1) = 1$ (provided that $n \leq \bar{n}$). However, when managers exert no effort or they supervise too many employees, coordination is extremely poor and $D(n,0) = n$. These assumptions guarantee that in the optimum all managers exert effort and that $2 \leq n \leq \bar{n}$,[26] but of course could be easily relaxed. Finally, I assume that the span of control is constant across layers, so that the hierarchy can be fully characterized by only two variables, n and L. This is important because it allows me to focus exclusively on the interaction between n and L, without having to worry, for instance, about what effect a change in the span of control at one layer might have on the span of control at another layer. Generalizing the model to consider interactions between variables at different layers would certainly be interesting but would also greatly complicate the analysis and is therefore left for future research.

A key element of this variant of the model is the presence of returns from specialization in information processing. Most of the information processing literature, with the important exception of Bolton and Dewatripont (1994), assumes that delegation is driven exclusively by the need to reduce delay. That has the unfortunate consequence that the number of levels in the organization tends to increase with urgency. This section illustrates how that effect can be offset when specialization is introduced into the model.

I model gains from specialization as follows. Let $\eta(w_L)$ be the time it takes a worker to process one unit of information. $\eta(w_L)$ is assumed to be increasing and convex in individual workload, w_L.[27] Thus, as

[25] Rajan and Wulf (2004) p.2. They quote this passage from a General Electric press release titled "GE Announces Reorganization of Financial Services; GE Capital to Become Four Separate Businesses", July, 26, 2002.

[26] For completeness, this claim is proven in the Appendix (Lemma A1).

[27] The definition of convexity adopted in this paper is the following. Let $X \subset \mathbb{R}$. A function $h : X \to \mathbb{R}$ is said to be

in Becker and Murphy (1992), smaller individual workloads and hence greater specialization allow workers to process each unit of information faster. Note that, realistically, returns from specialization are assumed to be decreasing.

Since all managers must exert effort in the optimum and the span of control is constant across layers, the problem of the hierarchy with specialization can be written as

$$\min_{n,L} T^S(n,L;\lambda) = \underbrace{N\eta\left(\frac{N}{n^{L-1}}\right) + \bar{\tau}\sum_{l=1}^{L-1} n^{l-1}}_{WB(n,L)} + \lambda\underbrace{\left[\frac{N}{n^{L-1}}\eta\left(\frac{N}{n^{L-1}}\right) + (L-1)(\bar{\tau}+b)\right]}_{DL(n,L)} \tag{11}$$

where without loss of generality $\bar{n} \geq n \geq 2$. Two features of (11) will be crucial in the following. The first is that now multiple layers can be optimal even when urgency is not important ($\lambda = 0$), since the organization may want the workers to specialize. This would never be the case in the basic model since there delegation is driven exclusively by the need to reduce delay. The second key property of (11) is that

Proposition 8 *$-T^S(n,L;\lambda)$ is supermodular in $(n,-L)$ on $S = \{(n,L) \in \mathbb{N}\times\mathbb{N} : \bar{n} \geq n \geq 2$ and $L \geq 2\}$.*

The notion of supermodularity provides a precise meaning to the intuition that the span of control and the number of levels are substitutes and indeed in this model their role is very similar, since they both allow the organization to increase parallel information processing by expanding its size. Interestingly, the result does not hold when $L = 1$: in that case, in fact, if the organization wants to increase the span of control, it must also necessarily create a new hierarchical level, and n and L are therefore complementary. The extreme case $L = 1$ can be ruled out by assuming, for instance, that $\bar{\tau} + \lambda[\bar{\tau} + b] < N\left[\eta(N) - \eta\left(\frac{N}{2}\right)\right] + \lambda N\left[\eta(N) - \frac{1}{2}\eta\left(\frac{N}{2}\right)\right]$, so that the one-worker hierarchy is dominated by a two-layer hierarchy with two subordinates. In the following, I will always assume that N is high enough that this condition holds.

To state the main result of this section, more definitions are needed. Fix n and let $S_L = \{L \in \mathbb{N} : L \geq 2\}$. Let $L_W^+(n)$ be the greatest $\arg\min_{L\in S_L} WB(n,L)$ and $L_D^-(n)$ be the smallest $\arg\min_{L\in S_L} DL(n,L)$ (these numbers clearly exist). I say that delegation is mainly driven by specialization, and write $L_W^+ \geq L_D^-$, if $L_W^+(n) \geq L_D^-(n)$ for all $\bar{n} \geq n \geq 2$.

Proposition 9 *Suppose that $L \geq 2$ and $L_W^+ \geq L_D^-$. Then the solution $(n^*, -L^*)$ to (11) is increasing in the strong set order in λ.*[28]

The intuition for this result is as follows. Provided delegation is mainly driven by specialization, as reducing delay becomes more important, the hierarchy tends to become flatter since the number of levels

(strictly) convex if $h(t'+c) - h(t+c) \geq (>) h(t') - h(t)$ whenever $t' > t$, $c > 0$, and the four points concerned lie in X.

[28] The statement of this proposition is somewhat imprecise, since program (11) may admit more than one solution. A more precise statement would be the following: Let $S' = \{(n,z) : (n,-z) \in S\}$, where S is defined as in Proposition 8. Then $\arg\min_{(n,z)\in S'} T^S(n,-z;\lambda)$ is increasing in the strong set order in λ.

that minimizes the overall objective function is too high from the point of view of minimizing delay.[29] Similarly, the span of control tends to increase because, as far as delay is concerned, larger spans are always beneficial. Finally, these two direct effects reinforce each other since n and L are substitutes. A slightly more general version of this result, which to the best of my knowledge is new in the monotone comparative statics literature, is stated and proved in the Appendix (Theorem A1).

The main problem with Proposition 9 is of course that it is not clear when one can reasonably expect delegation to be mainly driven by specialization. Intuition suggests that this condition is more likely to be fulfilled when $\bar{\tau}$ is low and b high, and when returns to specialization are substantial. The following example supports this conjecture:

Example 2 Suppose that $\eta(w) = \beta w^{\alpha}$, with $\beta > 0$, $\alpha \geq 1$. Then a sufficient condition for $L_W^+ \geq L_D^-$ is that $b \geq \frac{1}{2(2^{\alpha}-1)}\bar{\tau}$.

Condition $b \geq \frac{1}{2(2^{\alpha}-1)}\bar{\tau}$ is more likely to be satisfied when b and α are high (the latter being a measure of the strength of the gains from specialization), and $\bar{\tau}$ is low. The proof of example 2, which is included in the Appendix, also shows that $b > 0$ is a necessary condition for $L_W^+(n) \geq L_D^-(n)$. Thus, in the present framework, delegation can only be mainly driven by specialization if there is some source of delay in the communication process which is not accounted for in the wage bill.

6 Information Processing Hierarchies

This section focuses on a generalization of the production model which allows for the reporting of information to one's direct superior. It describes a situation where information must initially be processed by employees at the bottom of the organization (perhaps because they are the ones in direct contact with the customers) and then transmitted up the hierarchy through formal communication channels. The contribution of this section, besides checking the robustness of the main results of the paper, will be to provide a second rationale for the fact that superiors spend more time on coordination activities than their subordinates.

In information processing hierarchies, all the information is eventually transmitted to the top manager through reports which summarize and filter it. The basic model in Section 2 is extended as follows. As before, upper-level managers coordinate the work of their direct subordinates and bottom-level employees process raw information. However, now each subordinate is also assumed to send a report to his immediate superior, the size of which is proportional to the information he has processed. Formally, a manager at layer $l > 1$ processing an amount of information w_l will send a report of size $R_l = rw_l$ to his direct superior,

[29] Note in fact that both $WB(n, L)$ and $DL(n, L)$ are convex in L for fixed n, given our assumptions. This implies that the number of levels that minimizes (11) is between L_D^- and L_W^+.

where $r \geq 0$ is a parameter measuring information compression. Thus, since at layer L the workload of a bottom-level manager is $N \prod_{i=1}^{L-1} \frac{D(n_i,c_i)}{n_i}$, he will send a report of size $R_L = rN \prod_{i=1}^{L-1} \frac{D(n_i,c_i)}{n_i}$. Managers at layer $L-1$ receive their subordinates' reports $n_{L-1}R_L = rN \prod_{i=1}^{L-1} \frac{D(n_i,c_i)}{n_i}$, process all the 'relevant' information (see below) and send their own reports up the hierarchy until all the information has been communicated the top manager. Note that r must be sufficiently less than one because otherwise the organization would be better off if the superior processed that information alone.

Two important feature of this framework must be pointed out. The first feature is that skip-level reporting is not allowed: subordinates can only report to their direct superiors. This clearly involves some loss of generality since we know from Radner (1993) that skip-level reporting is an important feature of optimal information processing hierarchies. However, since this practice seems much less common in actual organizations than the information processing literature would suggest, it is probably useful to consider, at least as a first step, a scenario in which skip-level reporting is not allowed.

A second important feature of the model is that reports will in general contain duplications made a lower levels. For instance, the discussion above made clear that the reports received at layer $L-1$ contain duplications made at layer L. My analysis will distinguish two polar cases: the no-detection and the detection scenario. In the latter case, managers will be assumed to be able to detect and disregard duplications made at lower levels at no cost. In the former scenario, instead, this will be impossible and superiors will process duplications. Furthermore, their reports will also contain duplications made at lower levels. Needless to say, in reality we would expect something in-between these two polar scenarios. Finally, it must be stressed that both scenarios are generalizations of the basic model of Section 2 to the case where $r \neq 0$.

6.1 No-Detection Scenario

In the no-detection case, managers process and summarize the information they receive from their subordinates in a non-selective way. Thus, for example, since $R_l = rw_l$, the workload of a manager at layer $L-1$ is $w_{L-1} = rN \prod_{i=1}^{L-1} \frac{D(n_i,c_i)}{n_i}$. More generally, we have that

$$w_l = n_l R_{l+1} = n_l r w_{l+1} \tag{12}$$

for $l = 1, ..., L-1$ or, equivalently

$$w_l = r^{L-l} N \prod_{i=1}^{L-1} D(n_i, c_i) \left[\prod_{i=0}^{l-1} n_i \right]^{-1} \tag{13}$$

for all $l = 1, ..., L$. Clearly, equations (12) and (13) imply that higher-level managers process duplications made at lower levels, and also that their reports contain duplications. For all $L \geq 2$, the problem of the

hierarchy becomes

$$\min_{\mathbf{n},\mathbf{c}} \underbrace{\sum_{l=1}^{L}\left\{\eta N r^{L-l}\prod_{i=1}^{L-1}D(n_i,c_i)+\prod_{i=0}^{l-1}n_i\tau(c_l)\right\}}_{Wage\ Bill}+\lambda\underbrace{\sum_{l=1}^{L}\left(\eta N r^{L-l}\prod_{i=1}^{L-1}D(n_i,c_i)\left[\prod_{i=0}^{l-1}n_i\right]^{-1}+\tau(c_l)\right)}_{C(Delay)} \quad (14)$$

where of course $c_L = 0$ since bottom-level employees do not coordinate the work of other managers. Note that program (8) is a special case of (14) as the latter reduces to the former when $r = 0$. However, since $\prod_{i=1}^{L-1} D(n_i, c_i)$ still appears at all the levels of the hierarchy, it is easy to see that all the results of sections 3 and 4 carry over to the new scenario. In particular

Proposition 10 *Let the span of control be constant across layers and* $\mathbf{c}^* = (c_1^*, c_2^*, ..., c_{L-1}^*)$ *solve program (14). If* $\lambda \in \mathbb{R}_{++}$*, then* $c_1^* \geq c_2^* \geq ... \geq c_{L-1}^*$*. If* $\lambda = \infty$ *and* $-\frac{\tau'(c)}{D_c(n,c)}$ *is increasing in* c *for all* n*, then* $c_1^* = c_2^* = ... = c_{L-1}^*$.[30]

6.2 Detection Scenario

In this scenario, superiors are assumed to able to skip duplicated information at no cost (that is, without processing it) and therefore their reports contain no duplications. If superiors immediately detect duplicated tasks and do not process them twice, individual workloads are given by

$$w_l = r^{L-l} N \prod_{i=1}^{\min\{l,L-1\}} D(n,c_i)\left[\prod_{i=0}^{l-1}n_i\right]^{-1} \quad (15)$$

for all $l = 1, ..., L$. The problem of the hierarchy becomes

$$\min_{\mathbf{n},\mathbf{c}} \sum_{l=1}^{L}\left\{\eta N r^{L-l}\prod_{i=1}^{\min\{l,L-1\}}D(n_i,c_i)+\prod_{i=0}^{l-1}n_i\tau(c_l)\right\}+\lambda\sum_{l=1}^{L}\left(\eta N r^{L-l}\prod_{i=1}^{\min\{l,L-1\}}D(n_i,c_i)\left[\prod_{i=0}^{l-1}n_i\right]^{-1}+\tau(c_l)\right) \quad (16)$$

for $\lambda \in \mathbb{R}_{++}$ and

$$\min_{\mathbf{n},\mathbf{c}} \sum_{l=1}^{L}\left(\eta N r^{L-l}\prod_{i=1}^{\min\{l,L-1\}}D(n_i,c_i)\left[\prod_{i=0}^{l-1}n_i\right]^{-1}+\tau(c_l)\right) \quad (17)$$

for $\lambda = \infty$. The crucial difference between (16)-(17) and all the programs studied above is that now there is an asymmetry between duplications made at upper levels and those made at the lower levels, since the former affect a greater number of levels than the latter do. In particular, it should be clear that the organization now has an additional incentive to reduce duplications at the top than it had in the production or in the no-detecting scenarios. Thus, if, as it is intuitive, duplications increase as the span of control gets larger, higher-level managers will typically spend more time communicating and have more limited spans of control than their subordinates:

[30] The proof is analogous to those of Proposition 2 and 3 and is therefore omitted.

Proposition 11 *Suppose $D(n,c) \leq D(n+1,c)$ for all n and c. Let $L \geq 3$ and $(\mathbf{n}^*, \mathbf{c}^*)$ be a solution to (16). Then there are no k,$h \geq 1$ such that $c_k^* < c_{k+h}^*$ and $n_k^* > n_{k+h}^*$, $k+h \leq L-1$. Moreover, if $D(n,c)$ is log-supermodular, then $c_k^* \geq c_{k+h}^*$ and $n_k^* \leq n_{k+h}^*$.*

Note that, as in Proposition 6, a log-supermodularity assumption must be imposed on the duplication function to ensure that in the optimum $c_k \geq c_{k+h}$ and $n_k \leq n_{k+h}$. However, this condition is now even 'more sufficient' than before because of the greater incentives to invest in communication effort and reduce the span of control near the top of the hierarchy. These incentives can perhaps be best appreciated in uniform hierarchies (the proof is analogous to that of Proposition 2 and is therefore omitted).

Proposition 12 *Suppose the span of control is constant across layers and $\mathbf{c}^* = (c_1^*, c_2^*, ..., c_{L-1}^*)$ solves either program (16) or (17). Then $c_1^* \geq c_2^* \geq ... \geq c_{L-1}^*$.*

Relative to production hierarchies and the no-detection case, senior managers now spend more time on coordinating than their subordinates not only because it is less costly (as it was before), but also because their effort has an impact on a greater number of levels. Indeed, (15) makes clear that a manager's communication effort only affects the workloads of the managers below him. Thus, when comparing the effects of an increase in, say, c_1 and c_3, we must take into account that there are layers (in this case layers 2 and 3) in which an increase in c_1 reduces processing time but an increase in c_2 does not. This asymmetry is most evident when only delay matters. In fact, in production hierarchies, Proposition 3 guarantees that under some conditions communication efforts are equalized across layers. Now, in contrast, under the same conditions it may well be the case that $c_1^* > ... > c_{L-1}^*$, for instance if solutions are interior and $D(n,c) = n^{1-c}$ and $\tau(c) = \frac{c}{1-c}$. Of course, this does not imply that empowerment will not occur, as it is still true that senior managers exert more effort than their subordinates for cost reasons, and that effect disappears as minimizing delay becomes more and more important. Thus, the same forces at work in the production scenario also operate in information processing hierarchies.

7 Final Remarks

Coordination is an essential ingredient for survival and success in competitive environments, and ways to enhance it are a central concern for modern management. Liaison groups, teamwork, the use of intranets are just a few examples of working practices that more and more firms utilize to create and promote 'horizontal linkages' between interdependent organizational units, and constitute an important source of competitive advantage for companies such as Du Pont, Merck and Cisco. This paper has made a first attempt to provide a characterization of the optimal assignment of coordination responsibilities in a hierarchical organization. Nevertheless, many important issues have been neglected.

One major omission is that issues related to the optimal provision of incentives have been assumed away by imposing that all agents share the same objective. However, incentive problems have already received a lot of attention in the literature, so I will say no more here. A second limitation of this paper, which I have already mentioned, is that tasks are simply assumed to be divided evenly among subordinates. This assumption is very convenient in that it allows me to restrict attention to balanced hierarchies, which are both tractable and quite realistic, at least as a first approximation. However, fully endogenizing the division of tasks among subordinates would certainly be desirable. A third assumption that may be worth relaxing is the fact that, in my framework, tasks are homogeneous across the levels of the hierarchy. This may be done, for instance, by parametrizing the duplication functions in a way that reflects the greater complexity of the job as one goes up the hierarchy. In particular, in that scenario, I would expect the result that senior managers coordinate more than their subordinates to be strengthened, provided of course that the marginal returns to coordination rise with complexity.

Finally, I would like to emphasize that duplication of effort is certainly not the only problem that poor communication brings about. In real organizations, in fact, it is also often the case that some tasks are not carried out ('information loss') or that they are carried out in a way that makes it difficult to integrate them with other tasks ('poor integration'). In the remainder of this section, I will try to argue that these issues can, at least to a certain extent, be incorporated in the present framework.

Consider loss of useful information first. Many authors have noticed that the greater the number of levels in a system, the greater will be the probability of "noise" in functional connections among organizations units (e.g., Williamson (1967), Starbuck (1971)). Assume therefore that the exact specification of the tasks delegated to subordinates is subject to random shocks, or "misunderstandings". In particular, suppose that with some probability these misunderstandings prevent some of the items (or, in a continuous model, an interval of positive measure) from being processed. Clearly, if the cost of the resulting loss of information is high enough, it might be optimal to specify partially overlapping tasks so that the probability that a particular piece of information is not processed is minimized. Similarly, consider a scenario where the firm's employees (say 1 and 2) must integrate their tasks, X and Y, with another task, Z. Ideally, only one of the workers should be put in charge of Z (say 1); communication between 1 and 2 would then ensure that Y and Z are properly coordinated. In practice, however, the integration of Z with Y would probably be more difficult than with X, perhaps because of intrinsic difficulties in articulating tacit knowledge or because of a lack of incentives to help. Clearly, a radical way to solve the problem would be to assign task Z to both agents, so that each of them would be responsible for an integrated task, X+Z or Y+Z.

The discussion above suggests that when the costs of information loss or poor integration are very high, a certain amount of redundancy in task assignments may indeed be optimal, and indicates two possible ways in which the duplication function may be micro-founded. The present model could then be reinterpreted by saying that communication reduces misunderstandings. All results would straightforwardly carry over

to the new scenario, with reductions in amount of 'optimal' duplications now providing a simple way to quantify the benefits of fewer misunderstandings.

Appendix: Omitted Proofs

Proof of Proposition 1. Clearly, in the optimum $L < \infty$, otherwise delay would be infinite. To show that in the optimum $\bar{n} \geq n_l \geq 2$ for all l, assume instead that either $n_l = 1$ or $n_l > \bar{n}$ for some l. In either case, the size of a manager's task at layer l and $l+1$ is the same, and therefore no reduction in delay accrues. Clearly the organization would be better off by eliminating layer l, thus saving at least the fixed communication cost b. Hence, neither $n_l = 1$ nor $n_l > \bar{n}$ can be optimal. The existence of a solution for the communication efforts follows from Weierstrass theorem. Thus a solution to programs (7) and (6) exists. ■

Proof of Example 1. I first show that, if $D(n,c) = n^{1-c}$ and $\tau(c) = \frac{c}{1-c}$, then $c_l^* > c_{l+1}^*$ and $\partial\left(\frac{c_l^*}{c_{l+1}^*}\right)/\partial\lambda < 0$, $l = 1, ..., L-2$. Then I show that (8) is strictly convex. The case $\tau(c) = \alpha c^\beta$, $\alpha > 0$, $\beta > 1$ is similar and therefore omitted.

Assuming interior solutions, the first order conditions with respect to c_k and c_{k+1} yield $\left(n^{k-1}+\lambda\right)\left(1-c_k^*\right)^{-2} = (1+\lambda)\eta N n^{L-\sum_{i=1}^{L-1} c_i^*}$ and $\left(n^{k}+\lambda\right)\left(1-c_{k+1}^*\right)^{-2} = (1+\lambda)\eta N n^{L-\sum_{i=1}^{L-1} c_i^*}$. Since $n \geq 2$, $c_k^* > c_{k+1}^*$. Furthermore

$$\frac{1-c_k^*}{1-c_{k+1}^*} = \left(\frac{n^k+\lambda}{n^{k-1}+\lambda}\right)^{-1/2}$$

which implies that $\lim_{\lambda\to\infty} \frac{c_l^*}{c_{l+1}^*} = 1$. Since $\partial\left(\frac{1-c_k^*}{1-c_{k+1}^*}\right)/\partial\lambda < 0$ is strictly increasing in λ, $\frac{\partial c_k^*/\partial\lambda}{\partial c_{k+1}^*/\partial\lambda} < \frac{1-c_k^*}{1-c_{k+1}^*}$. Since $\frac{c_k^*}{c_{k+1}^*} > 1 > \frac{1-c_k^*}{1-c_{k+1}^*}$, we also have $\frac{\partial c_k^*/\partial\lambda}{\partial c_{k+1}^*/\partial\lambda} < \frac{c_k^*}{c_{k+1}^*}$. This implies that $\partial\left(\frac{c_k^*}{c_{k+1}^*}\right)/\partial\lambda < 0$.

To show that $T(\mathbf{c})$ is strictly convex in $\mathbf{c}$, compute its Hessian matrix $H_{T(\mathbf{c})} = \left[\frac{\partial^2 T(\mathbf{c})}{\partial c_i \partial c_j}\right]_{i,j=1,...,L-1}$ where $\frac{\partial^2 T(\mathbf{c})}{\partial c_i \partial c_j} = (1+\lambda)\eta N n^{L-\sum_{i=1}^{L-1} c_i} (\ln(n))^2 \equiv a > 0\ \forall i,j,\ i \neq j$ and $\frac{\partial^2 T(\mathbf{c})}{\partial c_k^2} = (1+\lambda)\eta N n^{L-\sum_{i=1}^{L-1} c_i} (\ln(n))^2 + \left(n^{k-1}+\lambda\right)\left(1-c_k\right)^{-2} \equiv a + b_k > 0$.

By applying elementary row and column operations, the Hessian matrix can be rewritten as

$$\tilde{H}_{T(\mathbf{c})} = \begin{bmatrix} b_1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & b_2 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & & \ddots & & & & \vdots \\ 0 & \cdots & 0 & b_k & 0 & \cdots & 0 \\ \vdots & & & & \ddots & & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & b_{L-2} & 0 \\ a & a & \cdots & \cdots & \cdots & a & x_{L-1}+b_{L-1} \end{bmatrix}$$

where $x_{L-1} = ab_{L-1}\sum_{i=1}^{L-1}\frac{1}{b_i} > 0$. Since the determinant of a lower-triangular matrix is simply the product of its diagonal entries, $\det(H_{T(\mathbf{c})}) = \det(\tilde{H}_{T(\mathbf{c})}) > 0$. The same reasoning can also be applied to all the leading principal minors of $H_{T(\mathbf{c})}$. (For instance, the determinant of the second order leading principal submatrix of $H_{T(\mathbf{c})}$ can be written as $b_1 \times \left[b_2 + ab_2\left(\frac{1}{b_1}+\frac{1}{b_2}\right)\right] > 0$.) Consequently, $H_{T(\mathbf{c})}$ is positive definite at every point, $T(\mathbf{c})$ is strictly convex in $\mathbf{c}$ and vector $\mathbf{c}^*$ that solve the first order conditions is the unique global minimum. ■

Proof of Proposition 4. Clearly, if $c_1^*(L) = 1$, there is nothing to prove. Consider therefore interior solutions: $0 < c_1^*(L) < 1$. By differentiating (8) with respect to c_1, compute the marginal benefit (MB) and the marginal cost (MC) of an increase in c_1 in the optimum for given L:

$$\left.\frac{\partial T(\mathbf{c})}{\partial c_1}\right|_{\mathbf{c}=\mathbf{c}^*} = 0 \quad \Leftrightarrow \quad \underbrace{(1+\lambda)\,\tau'(c_1^*)}_{MC(L)} \quad = \quad \underbrace{-D_c(n,c_1^*)\left(1+\frac{\lambda}{n^{L-1}}\right)\eta N\prod_{i=2}^{L-1}D(n,c_i^*)}_{MB(L)}.$$

Note that, if $MB(L) \to \infty$ as $L \to \infty$, then $c_1^*(L) \to 1$ since $\tau'(\cdot) > 0$. Assume by contradiction that $c_1^*(L)$ does not converge to 1 as L grows large. Then there exists a sequence $\{L_k\}_{k=1}^{\infty}$ such that $|c_1^*(L_k) - 1| \geq \delta > 0$ for all k. Since the domain of c, $[0,1]$, is compact, we can choose this sequence so that $c_1^*(L_k) \to \hat{c}$, $|\hat{c}-1| \geq \delta$ and hence $D(n,\hat{c}) > 1$. Note that $MB(L) \geq -D_c(n,c_1^*)\left(1+\frac{\lambda}{n^{L-1}}\right)\eta N\left[D(n,c_1^*)\right]^{L-2}$. But this last inequality implies $\lim_{k\to\infty} MB(L_k) = \infty$, and therefore $c_1^*(L) \to 1$, a contradiction. An analogous argument shows that $c_1^*(L) = 0$ cannot be optimal when L is very large since $\prod_{i=1}^{L-1}D(n,c_i^*)$ would then diverge. ■

Proof of Proposition 5. The proof is by contradiction. Let $(\tilde{\mathbf{n}},\tilde{\mathbf{c}}) = ((\tilde{n}_1,\tilde{c}_1),...,(\tilde{n}_{L-1},\tilde{c}_{L-1}))$ be a solution to (7) and assume that $\tilde{c}_k < \tilde{c}_{k+h}$ and $\tilde{n}_k \geq \tilde{n}_{k+h}$ for some k and h, $1 \leq k < k+h \leq L-1$ (the case where $\tilde{c}_k \leq \tilde{c}_{k+h}$ and $\tilde{n}_k > \tilde{n}_{k+h}$ is similar). Consider a vector $(\check{\mathbf{n}},\check{\mathbf{c}})$ whose elements are the same as in $(\tilde{\mathbf{n}},\tilde{\mathbf{c}})$ except that $(\tilde{n}_k,\tilde{c}_k)$ and $(\tilde{n}_{k+h},\tilde{c}_{k+h})$ have been interchanged. That is,

$$(\check{\mathbf{n}},\check{\mathbf{c}}) = (...,(\tilde{n}_{k-1},\tilde{c}_{k-1}),(\tilde{n}_{k+h},\tilde{c}_{k+h}),(\tilde{n}_{k+1},\tilde{c}_{k+1}),...,(\tilde{n}_{k+h-1},\tilde{c}_{k+h-1}),(\tilde{n}_k,\tilde{c}_k),(\tilde{n}_{k+h+1},\tilde{c}_{k+h+1}),...).$$

Note that

$$\eta N\prod_{i=1}^{L-1}D(n_i,c_i) + \lambda\left(\eta N\prod_{i=1}^{L-1}\tfrac{D(n_i,c_i)}{n_i} + \sum_{l=1}^{L-1}\tau(c_l)\right)\Bigg|_{(\mathbf{n},\mathbf{c})=(\tilde{\mathbf{n}},\tilde{\mathbf{c}})} =$$
$$\eta N\prod_{i=1}^{L-1}D(n_i,c_i) + \lambda\left(\eta N\prod_{i=1}^{L-1}\tfrac{D(n_i,c_i)}{n_i} + \sum_{l=1}^{L-1}\tau(c_l)\right)\Bigg|_{(\mathbf{n},\mathbf{c})=(\check{\mathbf{n}},\check{\mathbf{c}})}.$$

However $\sum_{l=1}^{L-1}\prod_{i=0}^{l-1}n_i\tau(c_l)\Big|_{(\mathbf{n},\mathbf{c})=(\tilde{\mathbf{n}},\tilde{\mathbf{c}})} > \sum_{l=1}^{L-1}\prod_{i=0}^{l-1}n_i\tau(c_l)\Big|_{(\mathbf{n},\mathbf{c})=(\check{\mathbf{n}},\check{\mathbf{c}})}$. In fact, the only terms in $\sum_{l=1}^{L-1}\prod_{i=0}^{l-1}n_i\tau(c_l)$

that matter for the comparison are the ones from k to $k+h$ included. For every $k < l < k+h$ we have $(\tilde{n}_k - \tilde{n}_{k+h}) \prod_{i=0, i\neq k}^{l-1} \tilde{n}_i \tau(\tilde{c}_l) \geq 0$. Moreover, $\tau(\tilde{c}_k) + \prod_{i=k}^{k+h-1} \tilde{n}_i \tau(\tilde{c}_{k+h}) > \tau(\tilde{c}_{k+h}) + \prod_{i=k+1}^{k+h} \tilde{n}_i \tau(\tilde{c}_k)$ since

$$\left(\tilde{n}_k \prod_{i=k+1}^{k+h-1} \tilde{n}_i - 1\right) \tau(\tilde{c}_{k+h}) > \left(\tilde{n}_{k+h} \prod_{i=k+1}^{k+h-1} \tilde{n}_i - 1\right) \tau(\tilde{c}_k).$$

Thus $T(\tilde{\mathbf{n}}, \tilde{\mathbf{c}}, L) > T(\breve{\mathbf{n}}, \breve{\mathbf{c}}, L)$ and $(\tilde{\mathbf{n}}, \tilde{\mathbf{c}})$ cannot be optimal. This yields a contradiction. ■

Proof of Proposition 6. The proof proceeds by contradiction. Let $(\tilde{\mathbf{n}}, \tilde{\mathbf{c}})$ be a candidate solution and consider the following three cases:

(a) $\tilde{c}_k < \tilde{c}_{k+h}$ and $\tilde{n}_k > \tilde{n}_{k+h}$. This is never optimal by Proposition 5.

(b) $\tilde{c}_k < \tilde{c}_{k+h}$ and $\tilde{n}_k \leq \tilde{n}_{k+h}$. Let $(\breve{\mathbf{n}}, \breve{\mathbf{c}}) = (\tilde{\mathbf{n}}, \tilde{\mathbf{c}})$ except that $\breve{c}_k = \tilde{c}_{k+h}$ and $\breve{c}_{k+h} = \tilde{c}_k$. $(\breve{\mathbf{n}}, \breve{\mathbf{c}})$ improves upon $(\tilde{\mathbf{n}}, \tilde{\mathbf{c}})$ if $T(\tilde{\mathbf{n}}, \tilde{\mathbf{c}}, L) > T(\breve{\mathbf{n}}, \breve{\mathbf{c}}, L)$. This inequality is clearly fulfilled provided

$$A\left[D(\tilde{n}_k, \tilde{c}_k) D(\tilde{n}_{k+h}, \tilde{c}_{k+h})\right] \geq A\left[D(\tilde{n}_k, \tilde{c}_{k+h}) D(\tilde{n}_{k+h}, \tilde{c}_k)\right] \tag{18}$$

where $A = \eta N \left[1 + \lambda \left(\prod_{i=1}^{L-1} \tilde{n}_i\right)^{-1}\right] \prod_{i\neq k, k+h} D(\tilde{n}_i, \tilde{c}_i)$ and

$$\prod_{i=0}^{k-1} \tilde{n}_i \tau(\tilde{c}_k) + \prod_{i=0}^{k+h-1} \tilde{n}_i \tau(\tilde{c}_{k+h}) > \prod_{i=0}^{k-1} \tilde{n}_i \tau(\tilde{c}_{k+h}) + \prod_{i=0}^{k+h-1} \tilde{n}_i \tau(\tilde{c}_k). \tag{19}$$

Under our assumptions, the second condition is always fulfilled. Condition (18) is fulfilled for all $\tilde{c}_k < \tilde{c}_{k+h}$ and $\tilde{n}_k \leq \tilde{n}_{k+h}$ if $D(n, c)$ is log-supermodular.

(c) $\tilde{c}_k \geq \tilde{c}_{k+h}$ and $\tilde{n}_k > \tilde{n}_{k+h}$. The proof is analogous to that of part b). Let $(\breve{\mathbf{n}}, \breve{\mathbf{c}}) = (\tilde{\mathbf{n}}, \tilde{\mathbf{c}})$ except that $\breve{n}_k = \tilde{n}_{k+h}$ and $\breve{n}_{k+h} = \tilde{n}_k$. A sufficient condition for $(\breve{\mathbf{n}}, \breve{\mathbf{c}})$ to improve upon $(\tilde{\mathbf{n}}, \tilde{\mathbf{c}})$ is that

$$A\left[D(\tilde{n}_k, \tilde{c}_k) D(\tilde{n}_{k+h}, \tilde{c}_{k+h})\right] \geq A\left[D(\tilde{n}_k, \tilde{c}_{k+h}) D(\tilde{n}_{k+h}, \tilde{c}_k)\right] \tag{20}$$

where A was defined above and

$$\tilde{n}_k \sum_{l=k}^{k+h} \prod_{\substack{i=0 \\ i\neq k}}^{l-1} \tilde{n}_i \tau(\tilde{c}_l) > \tilde{n}_{k+h} \sum_{l=k}^{k+h} \prod_{\substack{i=0 \\ i\neq k}}^{l-1} \tilde{n}_i \tau(\tilde{c}_l). \tag{21}$$

Condition (21) is satisfied under our assumptions. Condition (20) is satisfied for all $\tilde{c}_k \geq \tilde{c}_{k+h}$ and $\tilde{n}_k > \tilde{n}_{k+h}$ if $D(n, c)$ is log-supermodular.

Together, (a), (b) and (c) imply that log-supermodularity of $D(n, c)$ is sufficient to guarantee that in the optimum $c_k \geq c_{k+h}$ and $n_k \leq n_{k+h}$. ■

Proof of Proposition 7. Obvious from the symmetry of equation (6). ■

The following lemma proves a claim in Section 5.

Lemma A1 Consider program (7). Under the assumptions made in Section 5 (including those on $\eta(\cdot)$), $\bar{n} \geq n^* \geq 2$ and $c_l^* = 1$ for all $l \leq L-1$ and $L \geq 2$.

Proof. Suppose not. In particular, suppose that $c_l^* = 0$. This implies that the size of an individual task at layer l and $l+1$ are the same, since $D(n,0) = n$. But then it is optimal to eliminate layer l, to avoid at the very least the fixed cost b. Thus $c_l^* = 0$ cannot be optimal. That $\bar{n} \geq n^* \geq 2$ follows immediately from Propositions 1. ■

To show that $-T^S(n, L; \lambda)$ is supermodular in $(n, -L)$ on S, the following result is useful.

Lemma A2 Let $f : M \to \mathbb{R}$, $M = (x, y) \in \mathbb{N} \times \mathbb{N} : \underline{x} \leq x \leq \bar{x}, \underline{y} \leq y \leq \bar{y}\}$ ($\bar{x}$ and $\bar{y}$ being possibly infinite). If

$$f(x+1, y+1) + f(x, y) \geq f(x+1, y) + f(x, y+1)$$

for all $(x, y), (x+1, y+1) \in M$, then f is supermodular in (x, y) on M.[31]

Proof. Iterating the inequality in Lemma A2 yields

$$f(x+1, y) - f(x, y) \leq f(x+1, y+1) - f(x, y+1) \leq ... \leq f(x+1, y+z) - f(x, y+z) \leq ... \tag{22}$$

and similarly

$$f(x, y+1) - f(x, y) \leq f(x+1, y+1) - f(x+1, y) \leq ... \leq f(x+d, y+1) - f(x+d, y) \leq ... \tag{23}$$

for some $(x+d, y+z) \in M$. From (23), it follows that $f(x+d, y) - f(x, y) \leq f(x+d, y+1) - f(x, y+1)$. Using (22), after some iterations we get $f(x+d, y) - f(x, y) \leq f(x+d, y+z) - f(x, y+z)$. Thus

$$f(x+d, y+z) + f(x, y) \geq f(x+d, y) + f(x, y+z)$$

and f is supermodular in (x, y) on M. ■

Proof of Proposition 8. We need to show that T^S is supermodular in (n, L) on S. The first step is to show that $\frac{N}{n^{L-1}}$ is supermodular in (n, L) on S. By Lemma A2, it suffices to check that

$$\frac{1}{(n+1)^L} + \frac{1}{n^{L-1}} \geq \frac{1}{n^L} + \frac{1}{(n+1)^{L-1}} \tag{24}$$

which, after some manipulations, can be written as

$$L \geq \frac{\ln\left(\frac{n}{n-1}\right)}{\ln\left(\frac{n+1}{n}\right)} \equiv q(n).$$

[31] Lemma A2 can be thought of as variant for integer-valued choice variables of the 'marginal' rule stating that, for a twice continuously differentiable $f : X \to \mathbb{R}$, $X \subset \mathbb{R}^l$, supermodularity is equivalent to checking that $\frac{\partial^2 f}{\partial x_i \partial x_j} \geq 0$ for all $x \in X$ and all distinct $i, j = 1, ..., l$.

Clearly $1 < q(n) < 2$ for $n \geq 2$, since $\frac{n}{n-1} > \frac{n+1}{n}$ and $q < 2 \Leftrightarrow n^2 - n - 1 > 0$ are both verified for $n \geq 2$. (Alternatively, one can verify that $q(n)$ is decreasing in n and that $\ln(2)/\ln(3/2) \simeq 1.709$.) Thus $\frac{N}{n^{L-1}}$ is supermodular in (n, L) on S.

Next, suppose that $\frac{1}{n^L} \geq \frac{1}{(n+1)^{L-1}}$ (the other case is similar). Then $\frac{1}{n^{L-1}} - \frac{1}{n^L} \geq \frac{1}{(n+1)^{L-1}} - \frac{1}{(n+1)^L}$. Let $\frac{1}{(n+1)^{L-1}} - \frac{1}{(n+1)^L} = \Delta > 0$ and $\frac{1}{n^{L-1}} - \frac{1}{n^L} = \Delta + z$, $z \geq 0$. For any increasing and convex function K

$$\begin{aligned} K\left(\frac{1}{(n+1)^L}\right) + K\left(\frac{1}{n^{L-1}}\right) &\geq K\left(\Delta + z + \frac{1}{(n+1)^{L-1}}\right) + K\left(\frac{1}{(n+1)^{L-1}}\right) \\ &\geq K\left(\frac{1}{n^L}\right) + K\left(\frac{1}{(n+1)^{L-1}}\right) \end{aligned}$$

where the first inequality follows from the convexity of K and the second from the fact that K is increasing. Thus K is supermodular in (n, L) on S, and so are, in particular, $N\eta\left(\frac{N}{n^{L-1}}\right)$ and $\frac{N}{n^{L-1}}\eta\left(\frac{N}{n^{L-1}}\right)$.

To complete the proof, note that

$$\bar{\tau}\sum_{l=1}^{L}(n+1)^{l-1} + \bar{\tau}\sum_{l=1}^{L-1}n^{l-1} > \bar{\tau}\sum_{l=1}^{L}n^{l-1} + \bar{\tau}\sum_{l=1}^{L-1}(n+1)^{l-1} \iff (n+1)^{L-1} > n^{L-1}, \tag{25}$$

is always fulfilled. Thus also $\bar{\tau}\sum_{l=1}^{L-1}n^{l-1}$ and obviously $(L-1)(\bar{\tau}+b)$ are supermodular. Proposition 8 follows from the fact that sums of supermodular functions are supermodular. ■

To prove Proposition 9 and Example 2, the following results will be useful:

Lemma A3 Let $p : \mathbb{N} \times \mathbb{N} \to \mathbb{R}_+$. Assume that $\min_{y\in\mathbb{N}} p(x, y)$ has solution. For fixed x, let $y_p^+(x)$ and $y_p^-(x)$ denote, respectively, the greatest and the smallest $\arg\min_{y\in\mathbb{N}} p(x, y)$. Furthermore, for fixed x, suppose $p(x, y)$ is convex in y. Then (i) $p(x, y)$ is (strictly) increasing in y for all $y \geq y_p^-(x)$ $(y \geq y_p^+(x))$ and (ii) $p(x, y)$ is (strictly) decreasing in y for all $y \leq y_p^+(x)$ $(y \leq y_p^-(x))$.

Proof. We only prove the first part of the lemma; the proof of the second part is similar. By definition of $y_p^-(x)$,

$$p(x, y_p^-(x)) \leq p(x, y_p^-(x) + 1).$$

By convexity, for any $d \in \mathbb{N}$,

$$p(x, y_p^-(x) + d) - p(x, y_p^-(x) + d - 1) \leq p(x, y_p^-(x) + d + 1) - p(x, y_p^-(x) + d). \tag{26}$$

Now suppose that $p(x, y_p^-(x) + d - 1) \leq p(x, y_p^-(x) + d)$ for some $d \in \mathbb{N}$. From (26), it follows that $p(x, y_p^-(x) + d) \leq p(x, y_p^-(x) + d + 1)$. We conclude by induction that

$$p(x, y_p^-(x)) \leq p(x, y_p^-(x) + 1) \leq ... \leq p(x, y_p^-(x) + d) \leq ...$$

and therefore $p(x, y)$ is increasing in y for all $y \geq y_p^-(x)$.

To prove that $p(x,y)$ is strictly increasing in y for all $y \geq y_p^+(x)$, note that by definition, $p(x, y_p^+(x)) < p(x, y_p^+(x)+1)$. The result then follows by replacing the previous induction hypothesis with $p(x, y_p^+(x) + d-1) < p(x, y_p^+(x)+d)$ and noting that $0 < p(x, y_p^+(x)+d) - p(x, y_p^+(x)+d-1) \leq p(x, y_p^+(x)+d+1) - p(x, y_p^+(x)+d)$. ■

For future reference, this simple corollary of Lemma A3 is stated:

Corollary 1A Let p be convex in y for all x. The following two conditions are equivalent:

(i) $p(x, y+1) - p(x,y) \geq 0$

(ii) $y \geq y_p^-(x)$.

Proof. That (i) implies (ii) follows from the fact that p is strictly decreasing for $y \leq y_p^-(x)$. The reverse also follows from Lemma 3A, and specifically from the fact that p is increasing for $y \geq y_p^-(x)$. ■

Theorem A1 Let $\lambda \in \mathbb{R}_{++}$, $f, g : \mathbb{N} \times \mathbb{N} \to \mathbb{R}_+$ and $T(x,y;\lambda) = f(x,y) + \lambda g(x,y)$. Assume that $\min_{y \in \mathbb{N}} f(x,y)$ and $\min_{y \in \mathbb{N}} g(x,y)$ have solution. Let $y_f^+(x)$ and $y_g^-(x)$ be, respectively, the greatest $\arg\min_{y \in \mathbb{N}} f(x,y)$ and the smallest $\arg\min_{y \in \mathbb{N}} g(x,y)$. Suppose that **(A1)** f and g are supermodular in (x,y), **(A2)** f and g are convex in y for all x, **(A3)** $g(x,\cdot)$ is decreasing in x, and **(A4)** for all x, $y_f^+(x) \geq y_g^-(x)$. Let $V = \{(x,z) : (x,-z) \in \mathbb{N} \times \mathbb{N}\}$. Then $\arg\min_{(x,z) \in V} T(x,-z;\lambda)$ is increasing in the strong set order in λ.

Proof. The proof is divided into a number of steps.

Step 1 For fixed x, $\arg\min_{y \in \mathbb{N}} T(x,y;\lambda) \in \left[y_g^-(x), y_f^+(x)\right]$.

Proof. Suppose by contradiction that there exists a $\hat{y} \in \arg\min_{y \in \mathbb{N}} T(x,y;\lambda)$ such that $\hat{y} > y_f^+(x)$ (the case when $\hat{y} < y_g^-(x)$ is similar). Clearly $f(x,\hat{y}) > f(x, y_f^+(x))$ by definition of $y_f^+(x)$. That $g(x,\hat{y}) \geq g(x, y_f^+(x))$ for $\hat{y} > y_f^+(x) \geq y_g^-(x)$ follows from Lemma A3. Thus $f(x,\hat{y}) + \lambda g(x,\hat{y}) > f(x, y_f^+(x)) + g(x, y_f^+(x))$. But this implies $\hat{y} \notin \arg\min_{y \in \mathbb{N}} T(x,y;\lambda)$, a contradiction. ■

Step 2 Let $W = \left\{(x,y) \in \mathbb{N} \times \mathbb{N} : y \in \left[y_g^-(x), y_f^+(x)\right]\right\}$ and $V' = \{(x,z) : (x,-z) \in W\}$. Then V' is a lattice and

$$\arg\min_{(x,z) \in V} T(x,-z;\lambda) = \arg\min_{(x,z) \in V'} T(x,-z;\lambda) = \arg\max_{(x,z) \in V'} -T(x,-z;\lambda). \tag{27}$$

Proof. Note that $y_f^+(x)$ and $y_g^-(x)$ are both decreasing in x since $\min_{y \in \mathbb{N}} f(x,y) = \max_{y \in \mathbb{N}} -f(x,y)$ and $-f$ is submodular in (x,y), and similarly for g. Thus $-y_f^+(x)$ and $-y_g^-(x)$ are increasing in x and V' is a lattice.

The first equality in (27) follows from Step 1, which implies that $\arg\min_{(x,y) \in \mathbb{N} \times \mathbb{N}} T(x,y;\lambda) = \arg\min_{(x,y) \in W} T(x,y;\lambda)$, and the fact that $y = -z$. The second equality is obviously true. ■

Step 3 $-T(x,-z;\lambda)$ is supermodular in (x,z) on S' and satisfies increasing differences in (x,λ) and (z,λ) on S'.

Proof. The first part of the claim follows from the supermodularity of f and g (and hence T) in (x,y). $-T(x,-z;\lambda)$ satisfies increasing differences in (x,λ) since g is decreasing in x by assumption. Thus it suffices to check that, for all $(x,z) \in S'$ such that $(x'',z'') \geq (x',z')$, $-[f(x'',-z'')+\lambda g(x'',-z'')] + [f(x',-z')+\lambda g(x',-z')]$ is increasing in λ. This is true whenever $g(x',-z') - g(x'',-z'') \geq 0$ or, equivalently,

$$[g(x',-z') - g(x'',-z')] + [g(x'',-z') - g(x'',-z'')] \geq 0.$$

Clearly $g(x',-z') - g(x'',-z') \geq 0$ by A3. $g(x'',-z') - g(x'',-z'') \geq 0$ follows from the fact that $y_g^-(x) \leq -z'' \leq -z' \leq y_f^+(x)$ (since $(x,z) \in S'$) and g is increasing in y on $[y_g^-(x), y_f^+(x)]$. ■

Proof of the Theorem. Step 2 and 3 ensure that all the conditions of Theorem 5 in Milgrom and Shannon (1994) are fulfilled. The result then follows. ■

Proof of Proposition 9. To prove the result it suffices to check that the assumptions of Theorem 1A are met. The proof of Proposition 8 shows that WB and DL are supermodular in (n,L) on $S = \{(n,L) \in \mathbb{N} \times \mathbb{N} : \bar{n} \geq n \geq 2 \text{ and } L \geq 2\}$. It is also easy to check that, for fixed n, $WB(\cdot,L)$ and $DL(\cdot,L)$ are convex in L and $DL(n,L)$ is decreasing in n for all L. Finally, for all n, $L_W^+(n) \geq L_D^-(n)$ by assumption. ■

Proof of Example 2. Fix n. From Corollary 1A we know that

$$DL(n,L_W^+ + 1) - DL(n,L_W^+) = \bar{\tau} + b - \frac{N}{n^{L_W^+ - 1}}\left[\eta\left(\frac{N}{n^{L_W^+ - 1}}\right) - \frac{1}{n}\eta\left(\frac{N}{n^{L_W^+}}\right)\right] \geq 0 \tag{28}$$

implies $L_W^+(n) \geq L_D^-(n)$. Furthermore

$$\bar{\tau} n^{L_W^+ - 1} > N\left[\eta\left(\frac{N}{n^{L_W^+ - 1}}\right) - \eta\left(\frac{N}{n^{L_W^+}}\right)\right] \tag{29}$$

since $WB(n,L_W^+ + 1) - WB(n,L_W^+) > 0$ by definition. Multiplying the right-hand side of (28) by $n^{L_W^+ - 1} > 0$ and rearranging terms yield

$$\left[bn^{L_W - 1} - N\frac{n-1}{n}\eta\left(\frac{N}{n^{L_W^+}}\right)\right] + \bar{\tau} n^{L_W^+ - 1} - N\left[\eta\left(\frac{N}{n^{L_W^+ - 1}}\right) - \eta\left(\frac{N}{n^{L_W^+}}\right)\right] \geq 0$$

Thus a sufficient condition for $L_W^+(n) \geq L_D^-(n)$ is that

$$b - (n-1)\frac{N}{n^{L_W^+}}\eta\left(\frac{N}{n^{L_W^+}}\right) \geq 0. \tag{30}$$

We now use the fact that $\eta(w) = \beta w^\alpha$. (29) becomes

$$n^{L_W^+} > N\left[(n^\alpha - 1)\frac{\beta n}{\bar{\tau}}\right]^{\frac{1}{1+\alpha}} \equiv q(n). \tag{31}$$

Since (30) is increasing in $n^{L_W^+}$, we can replace $n^{L_W^+}$ with $q(n)$ in (31) and obtain, after some manipulations, the following sufficient condition for $L_W^+(n) \geq L_D^-(n)$:

$$b \geq \frac{(n-1)}{(n^\alpha - 1)n}\bar{\tau}.$$

The inequality in Example 2 then follows from the fact that the right-hand side of this equation is decreasing in n. ■

Proof of Proposition 11. The proof is similar to those of Proposition 5 and 6 and only a sketch is given. It is useful to divide total information processing costs (wages and delay) into three parts, namely those incurred a) at layers 1 to $k-1$, b) at layers k to $k+h-1$ and c) at all the remaining levels. Consider a putative solution where $n_k > n_{k+h}$ and $c_k < c_{k+h}$.

a) Neither (n_k, c_k) nor (n_{k+h}, c_{k+h}) has any impact on information processing costs at layers 1 to $k-1$. Thus swapping (n_k, c_k) for (n_{k+h}, c_{k+h}) has no impact on (16).

b) Only (n_k, c_k) has an impact on the information processing costs at layers at layers k to $k+h-1$. Clearly, it is beneficial to swap (n_k, c_k) for (n_{k+h}, c_{k+h}), since duplications are decreasing in n (and $c_k < c_{k+h}$).

c) Both (n_k, c_k) and (n_{k+h}, c_{k+h}) have an impact on information processing costs (and of course on total coordination costs). The same arguments used to prove Proposition 5 show that it is beneficial to swap (n_k, c_k) for (n_{k+h}, c_{k+h}).

It is therefore possible to improve on the putative solution. This yields a contradiction.

The proof of the second part of the proposition is analogous to the proof of Proposition 6. ■

References

[1] ATHEY, S., (2002), "Monotone Comparative Statics Under Uncertainty", *Quarterly Journal of Economics*, **117**, 187-223.

[2] BECKER, G.S. and MURPHY, K.M. (1992), "The Division of Labor, Coordination Costs and Knowledge", *Quarterly Journal of Economics*, **107**, 1137-1160.

[3] BEGGS, A.W. (2001), "Queues and Hierarchies", *Review of Economic Studies*, **68**, 297-322.

[4] BESANKO, D., DRANOVE, D. and SHANLEY, M. (2000), Economics of Strategy (New York: John Wiley & Sons).

[5] BOLTON, P. and DEWATRIPONT, M., (1994), "The Firm as a Communication Network", *Quarterly Journal of Economics*, **109**, 809-839.

[6] BOLTON, P. and FARRELL J. (1990), "Decentralization, Duplication, and Delay", *Journal of Political Economy*, **98**, 803-826.

[7] BUREAU OF THE CENSUS (1995), "First Findings from the EQW National Employer Survey", EQW Catalog Number: RE01.

[8] CALVO, C.A. and WELLISZ, S. (1978), "Supervision, Loss of Control and the Optimal Size of the Firm", *Journal of Political Economy*, **86**, 943-952.

[9] CALVO, C.A. and WELLISZ, S. (1979), "Hierarchy, Ability and Income Distribution", *Journal of Political Economy*, **87**, 991-1010.

[10] CHANDLER, A.D. (1990), Scale and Scope: the Dynamics of Industrial Capitalism (Cambridge, Massachussets: Harvard University Press).

[11] CHWE, M.S.Y. (1995), "Strategic Reliability of Communication Networks" (mimeo).

[12] DESSEIN, W. and SANTOS, T. (2003), "The Demand for Coordination" (Working Paper, Chicago Graduate Business School).

[13] GALBRAITH, J.R. (1977), Organization Design (Reading, Massachusetts: Addison-Wesley Publishing Company).

[14] GIFFORD, S. (1992), "Allocation of Entrepreneurial Attention", *Journal of Economic Behavior and Organization*, **19**, 265-284.

[15] GUETZKOW, H. (1981), "Communications in Organizations", in Nystrom P.C. and Starbuck W.H. (ed.), Handbook of Organizational Design (Oxford: Oxford University Press).

[16] HARRIS, M. and RAVIV, A. (2002), "Organization Design", Management Science, **48**, 852-865

[17] HART, O. and MOORE, J. (1999), "On the Design of Hierarchies: Coordination Versus Specialization" (Harvard Institute for Economic Research, Discussion Paper No. 1880).

[18] KEREN, M. and LEVHARI, D. (1979), "The Optimum Span of Control in a Pure Hierarchy", *Management Science*, **25**, 1162-1172.

[19] KEREN, M. and LEVHARI, D. (1983), "The Internal Organization of Firms and the Shape of Average Costs ", *Bell Journal of Economics*, **14**, 474-486.

[20] KEREN, M. and LEVHARI, D. (1989), "Decentralization, Aggregation, Control Loss and Costs in a Hierarchical Model of the Firm", *Journal of Economic Behavior and Organization*, **11**, 213-236.

[21] MARIN, D. and VERDIER, T. (2003), "Globalization and the New Enterprise", *Journal of the European Economic Association, Papers and Proceedings*, **1**, 337-344.

[22] MAHONEY T.A., JERDEE T.H and CARROLL S.J. (1965), "The Job(s) of Management", *Industrial Relations*, **4**, 97-110.

[23] MEAGHER, K.J. (2003), "Generalizing Incentives and Loss of Control in an Optimal Hierarchy: the Role of Information Technology", *Economic Letters*, **78**, 273-280.

[24] MILGROM, and SHANNON, C. (1994), "Monotone Comparative Statics", *Econometrica*, **62**, 1109-1146.

[25] PRAT A. (1997), "Hierarchies of Processors with Endogenous Capacity", *Journal of Economic Theory*, **77**, 214-222.

[26] QIAN, Y. (1994), "Incentives and Loss of Control in an Optimal Hierarchy", *Review of Economic Studies*, **61**, 527-544.

[27] RADNER, R. (1993), "The Organization of Decentralized Information Processing", *Econometrica*, **61**, 1109-1146.

[28] RAJAN, R.G. and WULF, J. (2004), "The Flattening Firm: Evidence from Panel Data on the Changing Nature of Corporate Hierarchies" (Working Paper, Chicago Graduate Business School).

[29] RICE, S.A. (1940), "The Role and Management of the Federal Statistical System", *American Political Science Review*, **34**, 481-488.

[30] ROSS, S.M. (1983), Introduction to Stochastic Dynamic Programming (New York, Academic Press).

[31] ROTEMBERG, J.J. (1999), "Process- versus Function-Based Hierarchies", *Journal of Economics & Management Strategy*, **8**, 453-487.

[32] SAYLES, L.R. (1964), Managerial Behavior (New York: McGraw-Hill).

[33] SIMON, H.A. (1976), Administrative Behavior (New York: Free Press).

[34] SIMON, H.A., SMITHBURG D.W. and THOMPSON V.A. (1950), Public Administration (New York: Knopf).

[35] STARBUCK, W.H. (1971), "Organization Growth and Development", 11-141, in STARBUCK, W.H. (ed.), Organizational Growth and Development (Harmondsworth: Penguin Books).

[36] URWICK, F.W. (1956), "The Manager's Span of Control", *Harvard Business Review,* **34**, 39-47.

[37] VAN ZANDT, T. (1995), "Continuous Approximations in the Study of Hierarchies", *Rand Journal of Economics,* ***26***, 575-590.

[38] VAN ZANDT, T. (2004), "Balancedness of Real-Time Hierarchical Resource Allocation" (CEPR Discussion Paper No. 4276).

[39] VAYANOS, D. (2003), "The Decentralization of Information Processing in the Presence of Interactions", *Review of Economic Studies*, **70**, 667-695.

[40] WILLIAMSON, O.E. (1967), "Hierarchical Control and Optimum Firm Size", *Journal of Political Economy*, **75**, 123-138.

[41] WILLIAMSON, O.E. (1975), Markets and Hierarchies: Analysis and Antitrust Implications (New York, Free Press).