The Firm as a Communication Network

Patrick Bolton; Mathias Dewatripont

The Quarterly Journal of Economics, Vol. 109, No. 4 (Nov., 1994), 809-839.

Stable URL:
http://links.jstor.org/sici?sici=0033-5533%28199411%29109%3A4%3C809%3ATFAACN%3E2.0.CO%3B2-A

The Quarterly Journal of Economics is currently published by The MIT Press.

THE
QUARTERLY JOURNAL
OF ECONOMICS

Vol. CIX November 1994 Issue 4

THE FIRM AS A COMMUNICATION NETWORK*

PATRICK BOLTON AND MATHIAS DEWATRIPONT

This paper analyzes how organizations can minimize costs of processing and communicating information. Communication is costly because it takes time for an agent to absorb new information sent by others. Agents can reduce this time by specializing in the processing of particular types of information. When these returns to specialization outweigh costs of communication, it is efficient for several agents to collaborate within a firm. It is shown that efficient networks involve centralization, that individuals delegate tasks to subordinates only if they are overloaded, and that the number of transits to the top tends to be equalized across individual information items.

I. INTRODUCTION

This paper is concerned with the organization of administration, clerical work, and production inside firms. The internal organization of firms is seen as a communication network that is designed to minimize both the costs of processing new information and the costs of communicating this information among its agents.

Several leading economists and economic historians of the modern corporation have recognized for some time the importance of information processing and communication costs for the firm's

*We are grateful to Masahiko Aoki, Drew Fudenberg, Gérard Genotte, Bengt Holmstrom, Nobuhiro Kiyotaki, Eric Maskin, Gérard Roland, Jean Tirole, and especially Marjorie Gassner, Timothy Van Zandt, and an anonymous referee for their suggestions. We have also benefited from comments by seminar participants at Tilburg University, University of Bruxelles; University of Leuven; University of Louvain; Stanford University; University of California, Davis; Harvard University; the Massachusetts Institute of Technology; the University of Chicago; University of Illinois, Chicago; Purdue University; University of Barcelona; Delta, Toulouse University, Oxford University; London School of Economics; and the University of Edinburgh. Part of this work was done at Studienzentrum Gerzensee. We wish to thank the "Pôle d'Attraction Interuniversitaire" program of the Belgian Government which has supported this work under grant 26.

internal organization. Thus, Chandler [1966] has argued that the multidivisional corporation (the most common form of internal organization in large modern firms) has been developed in order to respond to the growing problem of handling an ever increasing flow of information.[1]

Yet, most formal economic analysis of organizations has downplayed information processing and communication costs and focused mainly on issues related to individual incentives (see, for example, the survey by Holmstrom and Tirole [1989]). We believe that economists have shown so much interest in incentive issues inside organizations in part because a well-developed theoretical apparatus was available to analyze these issues. In comparison, the theoretical understanding of the processing and dissemination of information in the presence of communication costs is limited. Our paper is one among a few recent attempts to fill this gap.

Note that, while we are particularly interested in the theory of the firm, our model concerns organizations whose boundaries may be different from the legal boundaries of firms. Indeed, it could concern *subsets* of firms (production units, managerial units), as well as *some interfirm* relations (repeated subcontracting relations, for example). Moreover, it is relevant for bureaucracies and nonprofit organizations as well as for corporations.

The model we develop puts the organization in an environment in which a steady flow of information arrives over time that the firm may process. This flow of information is too large to be processed entirely by any group of agents. We call this a situation of *information overload.* The organization's problem, in this environment, is to design a fixed communication network to handle this flow of information most effectively.

An important dimension of our model is the idea of *returns to specialization* in processing. We assume that by repeatedly processing the same type of information item an agent can lower his unit time of processing that type of item. This is the main reason in our model why a group of several (specialized) agents want to work and process information as a team within the organization. Each of these agents handles a different type of information and the different pieces of information are aggregated through the commu-

1. To quote: "The basic reason for the success (of the multidivisional form) was simply that it clearly removed the executives responsible for the destiny of the entire enterprise from the more routine operational activities, and so gave them time, information, and even psychological commitment for long-term planning and appraisal" [Chandler 1966, pp. 382–83].

nication network. Eventually, one agent receives all the processed information and makes a decision based on all the available information. A central idea of our paper is that the benefits of greater specialization achieved by having more agents team up within the same organizations (each one handling more specialized information) are partly (and sometimes entirely) offset by the increased costs of communication within the enlarged group of agents. Hence, an important determinant of the form of efficient networks in our model is this *trade-off between specialization and communication.*[2] The more specialized agents are, the more communication is necessary to coordinate all these agents' activities, and therefore the larger and more sophisticated the organizations are within which these agents work. This point is broadly consistent with the evolution of corporations toward greater size and sophistication in the past 200 years.

In order to economize on overall communication costs, an efficient network must have a centralized design. As stressed by Arrow [1974], centralization avoids unnecessary duplication in communication and thus economizes on overall communication costs.[3] In Section III we show that efficient networks are centralized not only because a single agent receives all the processed information, but also because each agent sends his information to at most one other agent. Hence, efficient networks take a pyramidal form. There is a wide variety of pyramidal forms. One form that is often considered in the literature is the hierarchy in which each agent has an equal number of subordinates. Another pyramidal form that has also received a lot of attention is what we call the conveyor belt in which each agent (except the bottom agent) has only one subordinate. Both of these classic organizations can be efficient in our setup for some parameter values, but the situations in which they are efficient are quite specific. Our analysis, however, suggests that in most cases the efficient network is similar to either or a combination of these two structures. The former structure is usually seen as an efficient form of organizing clerical work and administration, the latter epitomizes the organization of mass production. It is remarkable that from the simple setup developed here, emphasizing only returns to specialization in processing and

2. This theme has also been addressed by Becker and Murphy [1992].

3. "Since transmission of information is costly in the sense of using resources, especially the time of the individuals, it is cheaper and more efficient to transmit all the pieces of information once to a central place than to disseminate each of them to everyone" [Arrow 1974, p. 68].

communication costs, should emerge two classic forms of organization that at first sight seem totally unrelated.

The general ideas discussed in this paper date back to at least before the second World War (see Robinson [1934] or Kaldor [1934] among others), but they have remained at the periphery of most treatments of the economics of organizations. We have been greatly influenced by major recent writings on these themes, in particular, Marschak and Radner [1972], Arrow [1974], Radner [1992, 1993], Van Zandt [1990], and Radner and Van Zandt [1992].

The work of Radner and Van Zandt is closely related to ours. They also emphasize information processing and communication costs and are concerned with the design of efficient communication networks. However, their approach differs from ours in that they stress different objectives for the organization, and they consider a communication technology that is a special case of ours. They are concerned with the objective of minimizing delay in processing a given batch of items, and they focus attention on how rotation of some agents with idle time between different groups processing different batches of items can accelerate the average processing time of a given batch of items. As they do not assume any returns to specialization in processing particular items, the benefits of rotating agents (to best use their idle time) can be substantial in their setting.

There have been a few other recent attempts at modeling the firm's internal organization as a communication network. Precursors to Radner and Van Zandt are Keren and Levhari [1979, 1983] and Beckmann [1960, 1983]. These papers, however, restrict attention to a small subset of all feasible networks, the set of pyramidal networks where, at each level of the hierarchy, each agent has the same number of subordinates.[4] The work of Sah and Stiglitz [1986] implicitly assumes that communication is costly and compares two modes of organization, hierarchies and polyarchies. A set of papers, comprising Crémer [1980], Aoki [1986], and Geanakoplos and Milgrom [1991], studies the efficient allocation of information-processing tasks in the presence of information-processing costs. However, these papers assume that communication is costless and therefore are not concerned with the design of a communication network.[5]

4. This restriction is also common in models of internal organization based on incentives (see, for example, Williamson [1967], Calvo and Wellisz [1978], and Qian [1994]).

5. Besides the work of Radner and Van Zandt, two other recent studies are closely related to ours, Marschak and Reichelstein [1987] and Wernerfelt [1993]. The former paper emphasizes, as we do in Section III, the idea that in an efficient

The remainder of the paper is organized in three sections. Section II describes the model and discusses the main assumptions. Section III gives the main results, and Section IV provides interpretations in relation to the management literature.

II. The Model

In this section we first present the assumptions underlying our analysis. We then detail two main forces that induce individuals to communicate, even though communication is costly: concern for delay (which a recent literature emphasizes), and returns to specialization (which we choose to stress). For each paradigm we analyze for the sake of illustration an example with few individuals, in order to abstract from issues of network design.

II.1. The Setup

We consider a model with continuous time and an infinite horizon without discounting. We are interested in the internal organization of a firm whose only activity, for our purposes, is to process valuable information about its environment. This firm is infinitely lived. Its sole objective is to maximize the flow return from processing information. In order to do the processing, this firm can have agents at a given opportunity cost from a large pool of labor composed of individuals of identical ability. The information about its environment available to the firm at any given time is described as follows. At each instant t, new information about the environment is available. We call the information at any date a cohort. Each cohort has M information items. All cohorts have the same value of information, $R > 0$. All items in any given cohort must be processed in order to extract the informational content of the cohort.[6] On the other hand, information gathered about one cohort is of no use for another cohort: all cohorts are thus informationally independent.

Processing any given item takes time. Let $\tau > 0$ denote the time it takes to process one item. Then a single agent can process

network the number of communication links cannot be reduced without slowing down the rate at which the organization processes new information. The latter paper highlights a trade-off which is related to our trade-off between specialization and coordination, namely the trade-off between coordination and delay: in the presence of information processing and communication costs, if a firm reacts more quickly to changes in its environment, it is at the expense of better coordination.

6. Alternatively, one could see this paper as the first step of a two-step optimization problem: (1) finding the least costly way to process a cohort of M items; and (2) optimizing over M, for a given benefit function $R(M)$.

new cohorts only every τM periods. Thus, the maximum flow return a single agent can get is $R/(\tau M)$.

If several agents process the same cohort, they can obtain the informational value of the cohort only if the items processed separately by the agents are all *communicated* to at least one of the agents. Communicating information also takes time. As in Radner [1993], we focus on *reading time.* Specifically, if agent i communicates m_i processed items to agent j, the total communication time is given by $C(m_i) = \tau(\lambda + a\, m_i)$, where $\tau > 0$, $\lambda > 0$, and $a > 0$ are constant parameters and $\lambda + a < 1$. To understand this communication cost function, normalize, for example, τ to 1: it then takes a fixed cost λ to connect agents i and j. Then a unit time a is required to read or hear each item. Of course, communication makes sense only if $\lambda + a < 1$; that is, reading a processed item takes less time than processing itself. Time, moreover, can be saved in communication by amalgamating items into a single condensed report. Specifically, if m items are communicated in this fashion, the unit time cost is $(\lambda + am)/m$, which is decreasing in m. The possibility of amalgamating single items into a condensed report provides an intuitive justification for the assumption that the communication technology exhibits increasing returns to scale. Note that our work generalizes the communication technology studied by Radner and Van Zandt, who consider only fixed communication costs ($\lambda > 0$ but $a = 0$).

When several agents are involved in processing any given cohort, the total time used to extract the informational value of the cohort is the total processing time τM plus the total communication time. If the firm's objective is to maximize the flow return from processing information, it must organize itself internally so as to minimize the total average time spent processing any given cohort. One way of reducing time spent on any given cohort is to try to reduce the time spent communicating. In this respect, the best possible outcome in the model outlined so far is to have only one agent involved in processing any given cohort, for then no communication between agents is even necessary.

Communication, however, may become efficient if performance depends on *delay* in processing cohorts, or on *returns to specialization,* which we now detail in turn. In our analysis, we shall restrict attention to networks fixed *once and for all,*[7] and

7. Indeed, we assume that the network is set up once and for all when the firm is founded and cannot be modified in the future. We justify this assumption on the grounds that the firm's environment is perfectly stationary and that setup costs of

assume that labor costs of the firm concern hours *actually worked,* that is, excluding idleness.[8]

II.2. Minimizing Delay

With one agent per cohort, the delay is τM. This delay could be cut substantially if, say, two agents were processing any given cohort—even though the two agents would spend additional time communicating. To see this, observe that each agent can process in parallel half the items in the cohort, $M/2$. When processing is completed, agent 1 sends a report to agent 2, so that the earliest time by which the cohort is processed starting at time $t = 0$, is $\tau(M/2 + \lambda + aM/2)$, which is less than τM. Parallel processing thus reduces delay. However, reducing delay implies decreasing throughput, since it requires communication.

In general, more than two agents may be required in order to minimize delay. Now, with three or more agents processing a given cohort, the question arises of which communication structure between these agents minimizes delay in communication? Not all communication networks involve the same delay as can be easily seen in the following example with three agents. Suppose that each agent processes a third of a cohort, and contrast the following two communication networks: in network 1, agent 1 communicates with agent 2 who, when he is done, communicates in turn with agent 3. In network 2, agent 1 communicates with agent 3, and agent 2 also communicates with agent 3 (see Figure I).

In network 1, total delay is

$$\tau\left(\frac{M}{3} + \lambda + a\frac{M}{3} + \lambda + a\frac{2M}{3}\right), \tag{1}$$

while in network 2, total delay is

$$\tau\left(\frac{M}{3} + 2\left(\lambda + a\frac{M}{3}\right)\right). \tag{2}$$

---

the network are large. Second, networks cannot be made cohort contingent. In other words, individuals' tasks are the same for all cohorts. The justification for this assumption is again that any change in the network is too costly to be worthwhile. In practice, firms may be organized so as to allow for some rotation in tasks across cohorts. While rotation of tasks is clearly an interesting dimension of the internal organization of firms, we believe that it is helpful as a first step to abstract from this aspect. Indeed, the design of networks with rotation is substantially more complex.

8. With two agents or more in a network, we shall see that it is likely that some agents are idle some of the time. We assume that the firm must only compensate agents for the actual time they spend working and that agents can use their idle time elsewhere. This is a convenient assumption that allows us to abstract from the somewhat peripheral issue of how the firm ought to use this idle time efficiently.

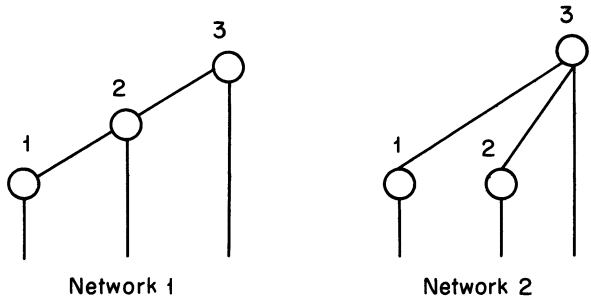

FIGURE I

Total delay is lower in network 2 than in network 1. The intuition for this result is again obvious. In network 1, delay is larger since there is duplication in communication: the items of agent 1 are communicated twice. The fact that some communication networks induce less delay than others has implications for organization design. Several authors, notably Radner [1993], Van Zandt [1990], and Wernerfelt [1993] have analyzed the internal organization of a firm in terms of delay minimization. These studies have been pretty successful in generating precise predictions about the efficient form of internal organizations.

However, while concerns for delay are undoubtedly empirically relevant, they do not appear to be of overriding importance in many settings. We choose here not to emphasize these concerns and to stress instead another dimension: *returns to specialization.*

II.3. Exploiting Returns to Specialization

In a well-known passage from *The Wealth of Nations,* Adam Smith argues that the gains from specialization arise from the repetition of the same task, which (1) improves dexterity, (2) saves time otherwise lost in switching from one activity to another, and (3) may lead to increased mechanization. These universally accepted principles also apply to information processing. Individuals do get better at repeatedly processing the same type of information. Time *is* lost in switching from processing one kind of information to another, and time *can* be saved with the help of computers in completing simple and mechanical processing tasks.

We model the gains from specialization as follows. We assume that each information item in any given cohort contains information of a specific type. Then, when an agent processes the same type

of item more frequently, the unit time costs of processing that item are lower. The same is assumed to be true for reading costs. More formally, a cohort can now be represented by an M-tuple $(n_1, \ldots, n_M)$. If an agent processes item n_i with frequency x, then his unit processing cost of that item is given by $\tau(x)$. Similarly, when he reads a report with m_i specific items with frequency x, it takes him $\tau(x)\ (\lambda + am_i)$ units of time to read it.[9] We assume that $\tau'(x) < 0$ and $\tau''(x) \geq 0$. For example, an organization such that only one agent processes any given cohort gives rise to a frequency of processing any given item of x_1, where x_1 is given by

$$x_1 = 1/[\tau(x_1)M]. \tag{3}$$

There exists at least one solution, x_1, to equation (3) if $\tau(0) < \infty$ and $\tau(\infty) > 0$. If there are multiple solutions, the relevant solution is obviously that with the highest value of x_1 (see Figure II).

The organization is interested in *minimizing total labor time spent per processed cohort.* In the one-agent case, this is $\tau(x_1)\ M$. In the presence of returns to specialization, the throughput of the organization may be increased if more than one agent is involved in processing any given cohort. To see this, consider once again the organization in which two agents process together any given processed cohort. Let m_1 be the number of items in any processed cohort handled by agent 1 and $M - m_1$ be the remaining items handled by agent 2. Suppose that communication takes the following form: for every processed cohort agent 1 sends his items to agent 2; the latter never sends his items to agent 1 (remember that throughout the paper we focus on fixed networks, without rotation of tasks).

With such a communication network the frequency of processing new cohorts is determined as follows: let x_2 be the endogenously determined rate at which new cohorts are processed. This rate depends on the allocation of the number of items to be processed in any given cohort to each agent. If m_1 is the number of items to be processed by agent 1, then his workload on any given cohort is $\tau(x_2)\ m_1$, and agent 2's workload is $\tau(x_2)\ (M - m_1 + \lambda + am_1)$. Suppose, for the sake of illustration, that an integer m_1 can be found such that the two agents' workloads are equal:

$$\tau(x_2)m_1 = \tau(x_2)[M - m_1 + \lambda + am_1]. \tag{4}$$

9. Alternatively, we could assume that gains from specialization do not apply to reading costs, but only to processing costs, without any fundamental changes in the results of this paper.

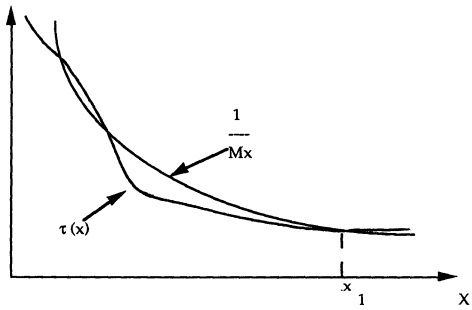

FIGURE II

Call m_1^* the solution to (4), so that $m_1^* = (M + \lambda)/(2 - a)$. Then x_2 is given by the solution to

$$x_2 = 1/[\tau(x_2)m_1^*]. \tag{5}$$

To see this, consider how the two agents can synchronize their activities so as to process a new cohort every time interval $\tau(x_2)\, m_1^*$: agent 1 processes m_1^* items in cohort t; after time $\tau(x_2)m_1^*$ has elapsed, he sends a message containing the m_1^* items of cohort t to agent 2 and immediately starts processing the new cohort, $t + 1$. Meanwhile, agent 2 spends time $\tau(x_2)\ (\lambda + am_1^*)$ reading agent 1's message on cohort $t - 1$; then agent 2 switches to processing $(M - m_1^*)$ items in cohort t; all this takes total time $\tau(x_2)\, m_1^*$. When agent 2 has completed the processing of cohort t, he turns to reading agent 1's report on cohort t and so on. In sum, when the two agents' workloads are equalized, they are both fully employed in processing cohorts at a rate of $x_2 = 1/(\tau(x_2)\, m_1^*)$. Total labor time per cohort is $\tau(x_2)\ (M + \lambda + am_1^*)$.

What happens when their workloads are not equal? Suppose first that $m_1 > m_1^*$. Now agent 1 can process cohorts only at a slower rate than x_2 since his workload is higher. The same is true for agent 2 for all the cohorts he works on with agent 1. But, when $m_1 > m_1^*$, agent 2's workload is less than agent 1's, so that agent 2 is idle for some of the time. As said before, we assume that agent 2 uses this idle time elsewhere, so that the relevant time cost for the firm is only the time spent in firm activities. Given this assumption, we can determine the efficiency of various allocations of items to agents by comparing the total average time spent on each cohort under any given allocation. Now, with $m_1 > m_1^*$, total labor time spent on each cohort is $\tau(\hat{x}_2)\ (M + \lambda + am_1)$, where $\hat{x}_2$ is the

endogenous frequency achieved when agent 1 processes m_1 items. Given that $m_1 > m_1^*$, we know that $\tau(\hat{x}_2) > \tau(x_2)$, so that total labor time per cohort is higher when $m_1 > m_1^*$. Thus, all allocations of items such that agent 2's workload is lower than agent 1's are inefficient.

What about the case $m_1 < m_1^*$? Agent 2 determines the frequency, which equals $\tilde{x}_2 = 1/(\tau(\tilde{x}_2)\ (M - m_1 + \lambda + am_1))$. Total labor time per cohort is $\tau(\tilde{x}_2)\ (M + \lambda + am_1)$. Two effects are now taking place in comparison with the equal workload frequency x_2: (1) since agent 2 works more, $\tilde{x}_2$ is lower than x_2; (2) agent 2 processes more items directly which means less communication costs for a given frequency. Effect (1) favors equal workloads, while effect (2) does not. The shape of $\tau(.)$ will determine whether it is optimal for loads to be equalized or not.

The above example captures the trade-off between returns to specialization due to high frequency and communication costs: it may pay for an agent to delegate part of the job to another agent in order to increase frequency, but delegation induces communication costs. The result that $m_1 \leq m_1^*$ reflects a general principle: delegation is minimized in that the receiver works at least as much as the sender in order to economize on communication costs. Overall, the shape of returns to specialization can lead to anything from no delegation ($m_1 = 0$) to maximum delegation ($m_1 = m_1^*$). This example thus shows how, in the presence of returns to specialization and in the absence of any concerns about delay, it may be efficient to have several agents process any given cohort *despite* the increased time cost in communication.

In general, to fully exploit gains from specialization, it may be efficient to have more than two agents per cohort. Then, as we pointed out earlier, the question arises as to which communication structure between these agents is best suited to exploit these gains from specialization. This question is the main focus of Section III.

III. Efficient Communication Networks

When there are large gains from specialization arising from the repetition of the same tasks over time, an important dimension of management is *control of throughput* (that is, the rate at which cohorts are processed). The frequency with which cohorts are processed can be increased—other things equal—by carefully designing the synchronization of tasks and communications between agents. But another way in which total labor time per cohort

can be decreased is by reducing communication costs *at a given frequency*. Subsections III.1 and III.2 are concerned with this first problem. The first subsection shows why efficient networks take a pyramidal form where each agent in the network reports to only one other agent. The second subsection shows how efficient networks involve specialization not only in the processing of items but also in the aggregation of reports. Finally, the third subsection is concerned with the optimal frequency of processing cohorts. It shows that, when the returns to specialization are large, efficient networks take a familiar structure: the network may look like a regular pyramidal hierarchy (integer constraints permitting), or in other circumstances, it may look like a conveyor belt.

III.1. Efficient Networks Are Pyramidal

The organization has completely processed a cohort only when at least one agent has absorbed all the items in the cohort. As we hinted at in the introduction, minimizing overall communication costs requires that at most one agent reads all information items. In other words, centralization of information in the hands of only one agent is one important way in which communication costs can be economized. We show below that this principle extends to all layers in the network; that is, at all layers it is (almost always) efficient for an agent to communicate his information items to only one agent. It is convenient to interpret the receiver of this information as the hierarchical superior of the agent, bearing in mind, though, that there is no relation of authority between agents here. To establish this claim, we need to introduce some definitions.[10]

DEFINITION 1: LAYER. All agents involved only in processing items are in layer 0. Agents who receive messages from other agents are in higher layers: an agent is said to be in layer h if he receives a message from at least one agent in layer $h - 1$, and no message from agents in higher layers.

DEFINITION 2: PYRAMIDAL NETWORK. A network is pyramidal (or forms a pyramid) if any agent in any layer $h = 0,1,2, \ldots, H - 1$ sends his items to only one other agent (where H is the highest layer).

10. We restrict attention, without loss of generality, to *acyclic* networks; that is, networks where no agent receives, directly or indirectly, information from any of his direct or indirect superiors.

We begin our analysis with an intuitively obvious result which, however, has sufficiently important implications that it is stated separately. In order to do this, note that each individual i in the network will have a workload per cohort of $\tau(x)$ $[y_i + n_i\lambda + aM_i]$, where x is the frequency of the network (defined by $x = (\max_i (y_i + n_i\lambda + aM_i))^{-1}$; y_i is the number of raw items processed by i; n_i is the number of reports read by i; and M_i is the total content of these n_i reports). Since subsections III.1 and III.2 take $\tau(x)$ as given, let us normalize it to 1 until the beginning of subsection III.3. We can then define $T^* = \max_i (y_i + n_i\lambda + aM_i)$ as the maximum time spent by an agent on any processed cohort.

PROPOSITION 1. In any efficient network, (1) an agent in layer $h \geq 2$ has a workload of at least $T^* - \lambda$, and (2) an agent in layer $h \geq 1$ has a workload of at least $T^* - 1 + a$.[11]

Proof of Proposition 1. Consider (1) first. Suppose by contradiction that one agent in layer $h \geq 2$ has a workload of less than $T^* - \lambda$. Let this be agent A_h. Given that agent A_h is in a layer at least as high as layer 2, he receives at least one message from an agent in layer $h - 1$, say from agent A_{h-1}. Now, agent A_{h-1} in turn receives messages from layer $h - 2$. Let m be the number of items in one of these messages.

If agent A_h has a workload of less than $T^* - \lambda$, he can receive at least one single message—that was transiting through agent A_{h-1}—directly from layer $h - 2$ without increasing his workload beyond T^*. To see this, note that agent A_h already receives some messages from layer $h - 2$ transiting through agent A_{h-1}. Take one of these messages, say message m, and have this message be sent directly to agent A_h, without transiting through agent A_{h-1}. Then agent A_h's workload increases by λ, but agent A_{h-1}'s workload decreases by $\lambda + am$. This simple reorganization of the network saves the organization a total time per cohort of am, without otherwise affecting the organization's performance. Thus, the new network is superior.

If after this reorganization agent A_h's workload is still below $T^* - \lambda$, the same operation can be repeated with another message transiting through agent A_{h-1}, and so on until agent A_h's workload is greater than or equal to $T^* - \lambda$.

Conceivably, all messages transiting through agent A_{h-1} may have been transferred directly to agent A_h, and yet agent A_h's

11. We thank Timothy Van Zandt for suggesting this second result.

workload is still less than $T^* - \lambda$. In that case further improvements can be obtained by transferring messages received by other agents in layer $h - 1$ who do not communicate with agent A_h.

All these local improvements may lead to the elimination of agents in layer $h - 1$ or conceivably to the elimination of the entire layer $h - 1$, or even of all intermediate layers, as long as agent A_h's workload is less than or equal to $T^* - \lambda$. In the extreme case where all intermediate layers are eliminated, agent A_h ends up in the top layer, which is then layer 1. This completes the proof of (1).

Part (2) can be proved similarly, as sketched now: allow the individual in layer $h \geq 1$ that has a workload lower than $T^* - 1 + a$ to directly process a raw item that was previously being processed by one of his own (direct or indirect) subordinates. This will increase his load by at most $1 - a$, and save the network at least a in total labor time per processed cohort.

QED

One important implication of Proposition 1 is that the *delegation* of tasks or items by an agent (at any layer of the network) to his subordinates only arises as a *result of that agent's work overload.* If the agent is not overloaded, he does not delegate items to subordinates. The strength of this proposition lies in the fact that this principle applies to all agents in the network.[12]

The basic insight behind Proposition 1 is straightforward: an efficient network minimizes the number of agents through which a given item transits averaged over all items per cohort. Thus, if an agent in layer $h \geq 2$ has time available, he should take on some of the load of his subordinates (that is, of the agents in lower layers with whom he communicates) so that the items taken on by the agent no longer transit through these subordinates. This simple insight is a powerful organizing principle of efficient networks.

A second important organizing principle is that—subject to all agents having a workload less than or equal to T^* and subject to the top agent getting all the information items—an efficient network minimizes the number of communication channels between agents. This requires in particular that—subject to integer constraints—any agent in the network send his items to at most one other agent (the unique direct hierarchical superior of the agent). It is obvious

12. This effect, also present in the work of Radner and Van Zandt, is a more basic explanation of delegation than the existing explanations based on incentive considerations where delegation is a commitment device to either influence third parties (as, for example, in Bonanno and Vickers [1988]) or to control ex post opportunism (as in Grossman and Hart [1986] or Aghion and Tirole [1993]).

that it is not efficient for an agent to send the same items to several other agents; in our setting this is just wasteful duplication. But the next proposition establishes the stronger result that in an efficient network an agent does not even send different items to several different agents. In other words, nonpyramidal networks are dominated by pyramidal networks (integer constraints permitting).

The basic logic of the argument behind the next proposition is that any network in which an agent has several direct hierarchical superiors can be turned into a pyramidal network by dividing up this agent's workload among as many new agents as there are direct hirarchical superiors, so that each new agent ends up having only one direct hierarchical superior. If this operation does not increase the total number of communication links, then a strict improvement can be obtained, since these new agents have spare time available which they may use to take up some of their subordinates' workload. We know from Proposition 1 that this involves a strict reduction in total communication time per cohort.

Denote the agent with multiple direct superiors as agent j. The division of agent j's workload into several smaller workloads may increase the total number of communication links if agent j's direct subordinates need to communicate with several new direct superiors as a result of the breakup in agent j's workload. Because of integer constraints we cannot rule out that one of his direct subordinates is forced to communicate with several of the new agents. Such integer problems, however, are unlikely to arise. We are unfortunately unable to say whether they ever arise at all. However, we can identify a large class of networks in which these integer problems are not present. These networks satisfy the following property.

PROPERTY (U). Let $\{m_j\}$ denote the collection of messages received by any agent j in the nonpyramidal network. Then, the set of messages that agent j sends to his direct superiors, $\{m^j\}$, is a partition of the set of messages received by agent j (that is, agent j does not disaggregate any message he receives).[13]

13. In many situations it is not even *feasible* to disaggregate messages received from subordinates. For example, if the message is a balance sheet summarizing all transactions in a given time interval, it may not be possible to recover the detail of all transactions from the balance sheet; in other words, it may not be feasible to disaggregate the balance sheet. Similarly, if the message is the sum of n numbers, it is not possible to recover exactly all n numbers from the sum. If it is not feasible to disaggregate messages for these reasons, then Property (U) is automatically satisfied.

PROPOSITION 2. All nonpyramidal networks satisfying Property (U) are dominated by some pyramidal network. They are strictly dominated when the nonpyramidal network has an agent with several direct superiors in any layer h higher than 1.

Proof of Proposition 2. Assume not. Then there exists at least one agent, say agent j, who has at least two superiors, agents k and l. Let m^k be the message sent to k and m^l the message sent to l. Consider the following transformation of this nonpyramidal hierarchy.

Step 1. Divide agent j's workload between agent j and a new agent $\hat{j}$. The former deals with workload m^k and the latter with workload m^l. Since Property (U) holds, this division of agent j's workload leaves the total number of communication links unchanged. Recall that we have assumed that agents are compensated only for the actual time they spend working for the organization. Therefore, adding agent $\hat{j}$ does not increase the organization's cost (total hours worked are unchanged), so that the new pyramidal network cannot be worse than the nonpyramidal network.

Step 2. Suppose that agent j is in layer $h > 1$. Once step 1 is completed, agents j and $\hat{j}$'s workloads are each strictly less than $T^* - \lambda$ (since they read at least one less message). At least one of these two agents has a subordinate in layer $h - 1$. Thus, by Proposition 1 the efficiency of the new pyramidal network can be improved, so that the nonpyramidal network is strictly dominated by a pyramidal network.[14]

QED

To close this subsection, we show how the principle of minimizing the number of communication links also leads to (almost) full employment in layer 0, thanks to another rule applying to all subordinates of a single superior.

PROPOSITION 3. Let $(j_1, \ldots, j_n)$ be a group of n agents communicating directly with the same superior. Then, in any efficient network, it is not feasible to divide the total workload of all n agents between $n - 1$ agents without violating the workload constraint T^*.

14. The reason why a nonpyramidal network with agents in layer $h = 1$ having multiple superiors may not necessarily be strictly improved upon is that when step 1 in the above transformation is completed, agents in layer 1 may not have enough idle time available to get involved in processing items.

Proof of Proposition 3. If it is feasible to divide the total load of the n agents among $n - 1$ of them, then a communication link can be eliminated, which would save the common superior at least λ. If such time savings are available, the network could not be efficient.

QED

Note that, while Propositions 1 and 2 focused on economizing on variable communication costs ($a > 0$), Proposition 3 concerns fixed communication costs ($\lambda > 0$).

To sum up, this subsection has identified two important organizing principles of efficient networks: (1) in any efficient network it should not be possible to reduce the number of communication links without otherwise affecting the performance of the organization; and (2) in any efficient network the average number of agents through which any given item transits is minimized. We have shown that these principles imply that efficient networks are likely to be pyramidal networks in which almost all agents are almost fully employed.

III.2. Efficient Networks Have Little Skip-Level Reporting

In this subsection we illustrate how the efficient organization of a pyramidal network necessarily involves little skip-level reporting. The term *skip-level reporting* refers to the practice in organizations whereby an agent in layer h sends reports to an agent in layer $h + L$, where $L \geq 2$.[15] That is, the agent in layer h skips one or several layers in reporting to his direct hierarchical superior. In practice, the internal organization of firms may allow for some skip-level reporting, but typically the extent of skip-level reporting seems to be limited.[16] We show in this subsection that skip-level reporting goes against the objective of minimizing the average number of transits any item must go through before reaching the top. Therefore, the extent of skip-level reporting in an efficient network is limited.

Consider a pyramidal network in which there is no skip-level reporting at all. In such a network any agent in layer $h \geq 1$ has direct subordinates only in layer $h - 1$. Since each agent in layers $h \geq 1$ has at least two subordinates, it follows immediately that in

15. According to Radner [1990] this is the terminology used at AT&T.

16. Direct evidence on the extent of skip-level reporting is hard to come by. Organizational charts do provide some indication of the structure of the internal organization, but these charts are often very sketchy and provide an artificially formal representation of hierarchical relationships between agents. Our knowledge of these matters should thus be seen as more than casual, to say the least.

such a network the size of reports received by any agent in layer h (as measured by the number of items contained in the report) is strictly increasing in h, the level of the layer. In other words, in a network without skip-level reporting, agents in higher layers handle more aggregated messages.

When there is skip-level reporting, this property is not necessarily satisfied. The most extreme form of skip-level reporting would have the agent in the top layer receive at least one message from one agent in layer 0. All the reports received by the top agent would then not be strictly greater than all reports received by any of his subordinates. In fact, some of them receive strictly greater reports than the smallest report the top gets (whenever the number of layers is greater than three). The next proposition establishes that such extreme forms of skip-level reporting are inefficient by showing that the smallest report received by any agent in an efficient network is greater than the largest report received by any of his direct or indirect subordinates. Thus, let $m_i(h)$ and $M_i(h)$ denote the size of, respectively, the smallest and the largest report *received* by agent i in layer h.

PROPOSITION 4. In any efficient pyramidal network, $m_i(h) \geq M_j(h - L)$, where j refers to any of agent i's direct or indirect subordinates in any layer $h - L$ ($L = 1, \ldots, h - 1$).

Proof of Proposition 4. If the proposition is true for direct subordinates, then it also holds for indirect subordinates, by transitivity. So, assume by contradiction that the proposition does not hold for direct subordinates. Then there exists one direct subordinate who *receives* at least one report of size $M_j(h - 1)$ such that $M_j(h - 1) > m_i(h)$. Let m_{ij} denote the size of the message *sent* by manager j to manager i.

We claim that there is a feasible improvement in the network where managers j and i swap reports: manager j now receives the message $m_i(h)$ and manager i receives the message $M_j(h - 1)$. This swap is feasible since the workload of manager i is unaffected and that of manager j is reduced. Indeed, after the swap manager i reads the message of size $M_j(h - 1)$ directly and receives a message of size $[m_{ij} - M_j (h - 1) + m_i(h)]$ from manager j. While, before the swap, manager i reads a message of size $m_i(h)$ directly and receives a message of size m_{ij} from manager j. Thus, before or after the swap manager i's workload remains unchanged.

Manager j's workload, on the other hand, has now been reduced by $(M_j(h - 1) - m_i(h))$, so that the swap leads to a net

positive reduction in total communication costs. This contradicts the assumed efficiency of the network.

QED

Proposition 4 is closely related to Proposition 1, since once again the objective here is to minimize the number of agents through which any given item transits and thus to economize on variable communication costs.[17] Proposition 4 indicates that there will be limits to skip-level reporting. The next proposition (which is a corollary of Proposition 4) shows that in any network in which all layer-0 agents have the same workload it is never efficient for a layer-0 agent to skip more than one layer.

PROPOSITION 5. In any efficient network where layer-0 agents have the same workload (call it y, where y is the number of items processed), these agents communicate with superiors in either layer 1 or 2.

Proof of Proposition 5. Suppose, by contradiction, that an agent in layer 0 communicates with a superior in layer $h \geq 3$. The size of his message is y (the number of items he has processed). But, the superior in layer $h \geq 3$ has at least one subordinate in layer 2, who receives at least one message of size strictly greater than y. This contradicts Proposition 4.

QED

Efficient networks thus have the property that agents in higher layers handle larger reports in which more items have been condensed. This property is consistent with descriptions of the capital budgeting process in large organizations where senior managers decide on larger investment projects (which usually combine several smaller projects) than junior managers who have discretion over only small-sized projects (see, for example, the description in Brealey and Myers [1991, pp. 261–69]). The next subsection establishes that skip-level reporting in efficient networks is a response to integer constraints. Indeed, in the absence of such constraints there is no skip-level reporting. This subsection also reveals, however, that situations in which integer constraints are not binding are the exception.

17. An important corollary of Proposition 4 is that the number of subordinates of an agent must be decreasing with the layer in which that agent is. This is consistent with observed practices of firms. See Section IV.

III.3. Regular Networks

The previous two subsections may give the misleading impression that efficient networks are rather complex. In this subsection we attempt to dispel this impression by showing that in some cases the efficient network takes a familiar form. We shall focus on two well-known structures, the *regular pyramidal network* and the *conveyor belt.*

DEFINITION 3: REGULAR PYRAMIDAL NETWORK. A regular pyramidal network is a pyramidal network in which an agent in layer h larger than zero spends no time processing raw items, only receives messages from subordinates in layer $h - 1$, and only sends messages to a single superior in layer $h + 1$. All agents in the same layer have the same number of subordinates, and the number of agents in any layer h is strictly decreasing in h.

DEFINITION 4: CONVEYOR BELT. A conveyor belt is a pyramidal network such that (1) all agents process the same number of items, and (2) there is only one agent in each layer.

Figure III depicts a conveyor belt. It is easy to see from the figure that this network resembles an assembly line where each agent (except the two agents at each end of the line) receives an aggregated message, adds his items to the message, and sends a more aggregated message to the next agent. One can think of this assembly line as the classic conveyor belt in automobile assembly plants if one interprets items as parts, a cohort as a single car, and a message as a partially assembled car (we elaborate on this interpretation in Section IV). While assembly lines are usually associated with the organization of production, our analysis suggests that such networks may also be well-suited to the organization of administration and clerical work.

The internal organization of large firms is often represented as a regular pyramidal network; indeed many corporations display organizational charts that look like regular pyramidal structures. In practice, of course, these networks are not as regular as they are made to appear, and examples of such regular networks are actually rather uncommon. One nice example we came across is one from the military—the organization of the armies of Genghis Khan in which the smallest unit was composed of ten men and each higher unit had a number of men $N = 10^x$, $(x = 2,3,4)$.[18]

18. Here is how Marshall [1993] describes the organization of the armies of Genghis Khan: "[New conscripts] would join their unit, which might be an

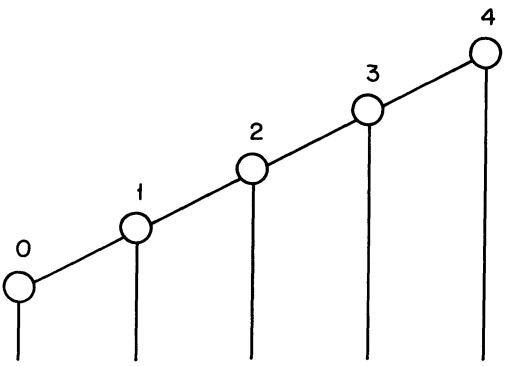

FIGURE III

It is not entirely surprising that real world examples of regular pyramidal networks (or, for that matter, of conveyor belt networks) are rare considering that many other variables besides those emphasized here affect the design of real organizations. Abstracting from all these variables, however, this subsection also suggests that the underlying architecture of internal administrative structures can be similar to a regular pyramid (and the organization of production may resemble a conveyor belt) yet this underlying structure may be obscured by integer constraints and the like. Integer constraints on the size of a cohort may only allow for regular pyramids in which some agents have a greater workload than others. As we know from Proposition 1, such a regular pyramid may then be dominated by another network with less overall delegation, in which the average number of transits any item goes through before reaching the top is lower.

Given that the regular pyramid and the conveyor belt are such different organizations, one might expect that they perform differently. We begin by addressing the question of when either of them is efficient and when is one likely to be better than the other.

It is convenient to first analyze this question in a special case. We shall assume that the communication technology is such that

arban—a simple unit of ten; a *jagun*—ten arbans or 100 men; a *minghan*—a regiment of ten jaguns or 1000 men; or a *tumen,* a division of ten minghans or 10,000 men" [p 37]. "The Mongols also employed an extremely effective and reliable system of signals, through flags, torches and riders who carried messages over great distances. This eventually provided them with one of the greatest advantages they ever took into the field: reliable and effective communications. It enabled all the Mongol units to remain in constant contact with each other and, through their remarkable corps of courriers, under control of a single commander" [p. 41].

the variable cost of communicating an additional item is zero ($a = 0$). In addition, we assume that the fixed cost λ and cohort size M, are such that $M = n^H = (1/\lambda)^H$, where n denotes the number of subordinates an agent has in the regular pyramidal network and H denotes the number of layers of the regular pyramid. This assumption ensures that the regular pyramidal network in which agents in layer 0 only process one item is such that agents at all layers have a full workload equal to $\tau(x^*)$. As layer-0 agents only process one item, the rate at which cohorts are processed by this network is $x^* = 1/\tau(x^*)$.

Notice that this network satisfies all the efficiency properties defined in the previous two subsections: the number of communication links cannot be reduced further without forcing at least one of the agents to take a workload greater than $\tau(x^*)$. In addition, minimizing the number of transits is not a concern here since the variable communication cost is zero ($a = 0$). However, this does not imply that the regular pyramid is efficient, since the total time it takes to process one cohort may still be reduced either by letting agents in layer 0 process more than one item or by letting agents in higher layers divide their time between processing and aggregating. In other words, the total time in processing a cohort may be reduced by forgoing some of the returns to specialization in order to save on total communication (or coordination) time. Thus, the efficiency of this network hinges on the efficiency of the rate of processing cohorts x^* attained under full specialization.

A sufficient condition for the regular pyramid with full specialization in processing to dominate any network where some agent processes at least two items is

$$\tau(y^*) \geq \tau(x^*)\left(1 + \lambda\left(1 + \frac{1}{n} + \frac{1}{n^2} + \cdots + \frac{1}{n^{H-1}}\right)\right), \tag{6}$$

where $\tau(y^*) = 1/(2y^*)$. Condition (6) says that the loss in returns to specialization incurred when at least one agent processes two items instead of one is greater than the total communication time per cohort in the regular pyramid with full specialization. Under (6) only two types of networks can be optimal: the regular pyramid with frequency x^* and the conveyor belt with frequency z^*, where $1/z^* = \tau(z^*)(1 + \lambda)$. This conveyor belt network allows for almost full specialization in processing, to the extent that all agents only process one item. But, as most agents share their time between processing an item and aggregating it with other items, the rate at which they process items is slightly lower than in the regular

pyramidal network. This loss in processing time, however, may be more than made up by the savings in total communication costs. Total communication costs under the conveyor belt are only $\tau(z^*)(M - 1)\lambda$, while under the regular pyramid they are $\tau(x^*)(M + n^{H-1} + n^{H-2} + \ldots + n)\ \lambda$. Thus, the comparison between these two regular networks brings out in particularly simple terms how the trade-off between the benefits of specialization and the costs of coordinating the tasks of specialized agents affects the choice of organizational form. The main proposition of this subsection provides a necessary and sufficient condition under which the conveyor belt is more efficient than the regular pyramid (with full specialization).

PROPOSITION 6. The conveyor belt with frequency z^* is more efficient than the regular pyramid with frequency x^* if and only if

$$\tau(z^*)\left(1 + \lambda\left(1 - \frac{1}{n^H}\right)\right) \leq \tau(x^*)\left(1 + \lambda\left(1 + \frac{1}{n} + \frac{1}{n^2} + \cdots + \frac{1}{n^{H-1}}\right)\right).$$

Proof. The total time spent on each cohort under the regular pyramid is given by

$$M(\tau(x^*)\left(1 + \lambda\left(1 + \frac{1}{n} + \frac{1}{n^2} + \cdots + \frac{1}{n^{H-1}}\right)\right). \tag{7}$$

The total time spent on each cohort under the conveyor belt is given by

$$M\tau(z^*)(1 + \lambda(M - 1)). \tag{8}$$

The comparison between (7) and (8) yields the proposition.[19]

QED

Once the number of agents in layer 0 is fixed, the conveyor belt is the network that minimizes the number of communication links. It is this property that makes it a candidate for efficiency. The

19. The proposition only compares the regular pyramid with the conveyor belt in the case of full specialization. However, as the notion of an item is a matter of convention, similar results can be established in other cases with less specialization (where agents in layer 0 process more than one item) by redefining the bundle of items processed by any agent as a single item.

above analysis, however, reveals that the minimization of communication links comes at the expense of returns to specialization.

When the communication technology is such that variable costs of communicating an item are positive ($a > 0$), there is an additional drawback to the conveyor belt: it does not minimize the average number of transits through which items must go before reaching the top. To see this, note that the workload of an agent is strictly increasing in the level of the layer. Now, if the difference in workloads between the top agent and the agent in layer 1 is such that $(1 + \lambda + aM) - (1 + \lambda + a) \geq \lambda$, then the conveyor belt can no longer be efficient since, by Proposition 1, the network can be organized more efficiently in the lower layers by having agents in those layers take on more than one subordinate. However, the efficient network in this case may still resemble a conveyor belt. The only difference between the efficient network and the conveyor belt in this case may be that the lower-layer agents have more than one subordinate in the efficient network.

Similarly, the regular pyramid is also likely to have drawbacks in the general setting. As the next proposition shows, the nature of integer constraints in the general setting is such that regular pyramids in which all agents have a full workload are unlikely to exist.

PROPOSITION 7. Regular pyramidal networks with more than three layers in which all agents have a full workload do not exist whenever $a > 0$.[20]

Proof of Proposition 7. See Appendix.

Integer problems are endemic in the general setting, and may prevent the optimality of regular pyramids in general. Once again, however, these integer problems should not hide the fact that the broad architecture of efficient networks in which the agents in the bottom layer are fully specialized in processing is likely to resemble a regular pyramidal network. Indeed, the results in subsections III.1 and III.2 point to this resemblance, since efficient networks are hierarchical and have little skip-level reporting.

The question remains whether regular pyramids in which all agents have a full workload are efficient when they exist. The next proposition establishes that such regular pyramids are undominated by any other network inducing the same frequency of

20. We thank Marjorie Gassner for giving us the proof of this result.

processing cohorts. It does so in the simple case where individuals in layer 0 only process a single item.

PROPOSITION 8. Consider regular pyramidal networks such that agents in layer 0 each process a single item and such that all agents have a full workload. Then, they strictly dominate all other networks with the same frequency of processing cohorts.

Proof of Proposition 8. See Appendix.

IV. INTERPRETATIONS AND CONCLUSIONS

We have shown that three broad principles determine the design of efficient communication networks: (1) the trade-off between specialization and coordination—the more specialized agents are the larger and more complex is the communication network coordinating agents' activities; (2) for a given level of specialization, an efficient communication network is such that the number of communication links between agents cannot be reduced without affecting the organization's performance (measured by the frequency with which new cohorts are processed); and (3) an efficient network is such that the average number of agents through which a given item must transit is minimized.

These three broad principles have far-reaching implications for the design of communication networks. First, in an efficient network each agent has at most one direct superior to whom he sends information. Second, an agent only delegates tasks to his subordinates when he is overloaded. It follows from these observations that efficient networks are pyramidal networks. Third, agents in higher layers handle longer reports and consequently have fewer subordinates than agents in lower layers. Finally, an efficient network resembles a regular pyramid when it is efficient to have agents fully specialized in either processing or aggregating. But, an efficient network may also resemble an assembly line when it is efficient for most agents to be involved in both processing and aggregating items. In sum, our setup is relevant both for the organization of administration and for the organization of production, since it can rationalize the types of structures one observes in both activities.

In the remainder of this section we briefly discuss how the results and insights of Section III relate to some of the main concerns of the management organization literature. An important part of that literature is based on direct observation of existing

internal administrative structures (see Mintzberg [1979]). Real world internal organizations must take into account many factors besides communication and processing costs; as a result, our approach cannot capture the richness of this descriptive analysis. However, communication costs and limited attention are major variables emphasized in the management literature.

IV.1. Span of Control and the Depth of the Organization

An important issue in the management literature is how wide should the span of control of any manager be and how many hierarchical layers should a firm have. Obviously, span of control and depth (measured by the number of layers) are closely interlinked. A wider span of control can lead to a flatter hierarchy and vice versa. The early literature on this subject has gone so far as to quantify the optimal number of subordinates per manager at each level.[21] Most discussions on depth of the hierarchy and span of control assume that the firm's internal organization takes the form of a regular pyramid. Why this is an efficient form is rarely considered.

In this respect, our analysis here complements these studies by providing one explanation (based on communication costs and limited attention) for the efficiency of regular pyramids. Our results suggest that the workload of all managers should be equalized, which implies that managers in higher layers have fewer subordinates than managers in lower layers since they read longer reports. This is consistent with the above observations about the smaller span of control in higher layers. Yet, our results also suggest that very regular pyramids are unlikely to be efficient in most environments and that as a result the important issue of the

21. In a survey by Koontz [1966], for example, a leading scholar—Urwick [1956]—is quoted as saying that "no superior can supervise directly the work of more than five or, at most, six subordinates whose work interlocks." Other scholars have suggested that the span should be smaller at the top (from 3 to 7) than at the bottom (from 20 to 30) of the hierarchy. Koontz also cites two empirical studies of the internal organization of more than 600 plants and companies in which over 80 percent of the firms had top executives with less than nine immediate subordinates. An interesting and rare experimental study by Carzo and Yanouzas [1969] compares two regular pyramids; both have 15 managers, one organization is flat and has only two layers (the boss and his 14 subordinates), the other is deep and has four layers (8 managers at the bottom and a span of 2 subordinates at each higher layer). Each organization is asked to solve a problem of supply allocation to different geographical market areas with varying demand. The results of this study show no significant difference between the two organizations in the average time taken to make the supply allocation decisions; however, overall coordination was better in the deep structure.

optimal span of control has perhaps been confined within an excessively narrow framework.

The management literature stresses other dimensions besides work-overload and communication costs, such as motivation of managers, making lower level employees more responsible and eager to take initiatives, etc. A natural avenue for future research is to include some of these dimensions into our model and explore how they affect the design of internal organizations.

IV.2. The Conveyor Belt and the Organization of Production

The conveyor belt has been at the center of most discussions on the organization of mass production ever since Taylor's [1911] seminal work on scientific management and the successful implementation of his ideas in many manufacturing firms. Perhaps the most famous example of a conveyor-belt organization is Henry Ford's automobile assembly plants. As we mentioned in Section III, we can interpret our model as a model of production. An item is then interpreted as an individual part, processing an item means fitting or producing a part that can be assembled into the emerging product, and aggregating items means assembling the fitted part. Finally, a cohort is interpreted as a fully assembled product. Accordingly, our model can be seen as one attempt at formalizing some of Taylor's ideas on the organization of production. Indeed, much of his work emphasizes the productivity gains from performing the same simple tasks repeatedly and the importance of synchronizing workers' activities so as to allow each worker to repeat the same task at a constant frequency (or to allow the introduction of automation).

Of course, in practice the whole process of producing a car or a television set is much more complex. Assembling parts is only one stage; other stages include the manufacturing and shipping of parts to the assembly plant, the training of workers, quality control, inventory management, etc. A complete model of the organization of production ought to include these stages.[22]

IV.3. Lower Communication Costs Lead to Flatter Hierarchies

A straightforward prediction of our model is that a reduction in communication costs (say, as a result of the introduction of the

22. Including them in a model like ours seems to be a very interesting avenue for future research. All the more so since many well-established production organizations seem to be currently undergoing profound changes. (see, for example, Womack et al. [1991], on the Japanese influence in car production).

telephone or computers) leads to a flatter and smaller organization. More precisely, a reduction in λ or a allows all agents to take on more items. Since agents can thus increase their workload, fewer agents are required to process a given number of items. Also, by Proposition 1, agents in layers $h > 1$ then take up more work from their subordinates until they are again fully employed. This process can only lead to a reduction in the number of layers.

Interestingly, there seems to be some empirical evidence that the computerization of firms has had the effect of reducing the number of layers inside firms (see Brynjolfsson et al. [1989] and Hagström [1991]). This evolution is sometimes interpreted as a move toward greater decentralization. According to our model, however, this move can also be interpreted as an increase in centralization, senior management increasing its span of control.

APPENDIX

Proof of Proposition 7. Call n_t the number of subordinates of layer t. Then, no idleness implies that $\tau(x^*) = (\lambda + a)n_1 = (\lambda + an_1)n_2 = \ldots = (\lambda + an_1 n_2 \ldots n_{t-1})n_t$ for $n_1, n_2, \ldots, n_t \in N$. For $a > 0$, we must have $n_1 > n_2 > \ldots > n_t$. Moreover, pyramids can be efficient only for $n_t \geq 2$. Of course, with only three layers, the above equations are simply $\tau(x^*) = (\lambda + a)n_1 = (\lambda + an_1)n_2$. Any $n_1 > n_2 \geq 2$ will generate $\lambda > 0, a > 0$, so that many solutions exist. What if $t > 2$? $\tau(x^*)$ is a normalization, so we can set it equal to 1. For $t > 2$, we must thus have at least

$$1 = (\lambda + a)n_1 \tag{7}$$

$$1 = (\lambda + an_1)n_2 \tag{8}$$

$$1 = (\lambda + an_1n_2)n_3. \tag{9}$$

Let us eliminate λ and a:

$$(7)\text{–}(8) \Rightarrow \lambda(n_1 - n_2) = an_1(n_2 - 1)$$

$$(8)\text{–}(9) \Rightarrow \lambda(n_2 - n_3) = an_1n_2(n_3 - 1).$$

Dividing and rearranging yields

$$n_1 = n_2 + \frac{(n_2 - 1)(n_2 - n_3)}{n_2(n_3 - 1)}. \tag{10}$$

The question becomes do there exist n_2, n_3, integers bigger than 1, such that the second term of the right-hand side of (10) is

an integer. In fact, no. First, notice that $(n_2 - 1)/n_2$ cannot be simplified at all. Indeed, otherwise one would have $n_2 - 1 = \alpha k_1$, and $n_2 = \alpha k_2$, with α being an integer greater than 1, and k_1 and k_2 being two integers. Then, we would have $n_2 - (n_2 - 1) = 1 = \alpha (k_2 - k_1)$. But this is a contradiction, since $\alpha > 1$ and $k_2 \geq k_1$. Consequently, n_2 cannot be simplified at all with $(n_2 - 1)$. And it cannot "disappear," that is, be fully simplified, with $(n_2 - n_3)$, since $n_2 - n_3 < n_2 - 1$. Thus, if n_2 and n_3 are integers, n_1 cannot be.

QED

Proof of Proposition 8. By Proposition 7 we can restrict attention to regular pyramids with at most three layers. Of course, a regular pyramid with two layers and no idleness is optimal. Consider thus a three-layer regular pyramid with frequency x^* such that $\tau(x^*) = 1/x^*$, since agents in layer 0 only process a single item. The size of the cohort is $n_1 n_2$, where n_i is the number of subordinates per individual in layer i, and $\tau(x^*) = n_1 (\lambda + a) = n_2 (\lambda + a n_1)$. On top of the $n_1 n_2$ agents in layer 0, the network employs $n_2 + 1$ individuals, and has communication costs equal to $(n_2 + 1) \tau(x^*)$. Is this optimal? It is certainly optimal for the individual at the top of the hierarchy to have n_2 subordinates. Sharing their workloads equally allows them all to be in layer 1. Not doing so means that at least one subordinate would be in layer 2 or higher, so that the head of the hierarchy would be in layer 3 or higher. But, then, by Proposition 5, he would not communicate at all with individuals in layer 0. Consequently, the network would have *more* than $(n_2 + 1)$ individuals above layer 0, and each information item would have to transit by *at least* as many individuals as in the regular case. Consequently, moving away from the regular pyramid *strictly* increases communication costs.

QED

LONDON SCHOOL OF ECONOMICS, CEPR, AND ECARE
ECARE (UNIVERSITE LIBRE DE BRUXELLES), CEPR, AND CORE

REFERENCES

Aghion, P., and J. Tirole, "Formal and Real Authority in Organizations," mimeo, IDEI and Oxford, 1993.

Aoki, M., "Horizontal vs. Vertical Information Structure of the Firm," *American Economic Review*, LXXVI (1986), 971–83.

———, *Information, Incentives and Bargaining in the Japanese Economy* (Cambridge: Cambridge University Press, 1988).

Arrow, K. J., *The Limits of Organization* (New York, NY: Norton, 1974).

Becker, G., and K. Murphy, "The Division of Labor, Coordination Costs, and Knowledge," *Quarterly Journal of Economics,* CVII (1992), 1137–60.
Beckmann, M., "Some Aspects of Returns to Scale in Business Administration," *Quarterly Journal of Economics,* LXXIV (1960), 464–71.
____, *Tinbergen Lectures on Organization Theory* (Heidelberg: Springer-Verlag, 1983).
Bonanno, G., and J. Vickers, "Vertical Separation," *Journal of Industrial Economics,* XXXVI (1988), 257–65.
Brealey, R., and S. Myers, *Principles of Corporate Finance* (New York, NY: McGraw-Hill, 1991).
Brynjolfsson, E., T. W. Malone, V. Gurbaxani, and A. Kambil, "Does Information Technology Lead to Smaller Firms?" mimeo, Center for Coordination Science, MIT, 1989.
Calvo, G., and S. Wellisz, "Supervision, Loss of Control, and the Optimal Size of the Firm," *Journal of Political Economy,* LXXXVI (1978), 943–52.
Carzo, R., and J. N. Yanouzas, "Effects of Flat and Tall Organization Structure," *Administrative Science Quarterly,* XIV (1969), 178–91.
Chandler, A., *Strategy and Structure* (New York, NY: Doubleday, 1966).
Crémer, J., "A Partial Theory of the Optimal Organization of Bureaucracy," *Bell Journal of Economics,* XI (1980), 683–93.
Geanakoplos, J., and P. Milgrom, "A Theory of Hierarchies Based on Limited Managerial Attention," *Journal of the Japanese and International Economies,* V (1991), 205–225.
Grossman, S., and O. Hart, "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration," *Journal of Political Economy,* XCIV (1986), 691–719.
Hagström, P., "The "Wired" Multinational Corporation: The Role of Information Systems for Structural Change in Complex Organizations," Ph.D. thesis, Stockholm School of Economics, 1991.
Holmstrom, B., and J. Tirole, "The Theory of the Firm," in *Handbook of Industrial Organization,* Vol. 1, R. Schmalensee and R. Willig, eds. (Amsterdam: North-Holland, 1989).
Kaldor, N., "The Equilibrium of the Firm," *Economic Journal,* XLIV (1934), 70–71.
Keren, M., and D. Levhari, "The Optimum Span of Control in a Pure Hierarchy," *Management Science,* XL (1979), 1162–72.
Keren, M., and D. Levhari, "The Internal Organization of the Firm and the Shape of Average Costs," *Bell Journal of Economics,* XIV (1983), 474–86.
Koontz, H., "Making Theory Operational: The Span of Management," *Journal of Management Studies,* III (1966); reprinted in H. Koontz, C. O'Donnell, and H. Weihrich, eds., *Management: A Book of Readings* (New York, NY: McGraw-Hill, 1980), pp. 232–40.
Marschak, J., and R. Radner, *Economic Theory of Teams* (New Haven, CT: Yale University Press, 1972).
Marschak, T. A., and S. Reichelstein, "Network Mechanisms, Informational Efficiency and the Role of Hierarchies," mimeo, Stanford Graduate School of Business, 1987.
Marshall, R., *Storm from the East: From Ghengis Khan to Khubilai Khan* (London: BBC Books, 1993).
Mintzberg, H., *The Structuring of Organizations* (London: Prentice Hall, 1979).
Qian, Y., "Incentives and Loss of Control in an Optimal Hierarchy," *Review of Economic Studies,* LXI (1994), 527–44.
Radner, R., "Hierarchy: The Economics of Managing," *Journal of Economic Literature,* XXX (1992), 1382–1415.
____, "The Organization of Decentralized Information Processing," *Econometrica,* LXI (1993), 1109–46.
Radner, R., and T. Van Zandt, "Information Processing in Firms and Returns to Scale," in *Annales d'Economie et de Statistique,* XXV–XXVI (1992), 265–98.
Robinson, E. A. G., "The Problem of Management and the Size of the Firm," *Economic Journal,* XLIV (1934), 240–54.
Sah, R. K., and J. Stiglitz, "The Architecture of Economic Systems: Hierarchies and Polyarchies," *American Economic Review,* LXXVI (1986), 716–27.

Taylor, F. W., *Principles of Scientific Management* (New York, NY: Harper & Row, 1911).
Urwick, L. F., "The Manager's Span of Control," *Harvard Business Review,* XXXIV, (1956), 39–47.
Van Zandt, T., "Efficient Parallel Addition," Bell Laboratories Discussion Paper, 1990.
Wernerfelt, B., "The Structure and Scope of Firms and Economies of Scale in Contracting," mimeo, MIT, 1993.
Williamson, O., "Hierarchical Control and Optimum Firm Size," *Journal of Political Economy,* LXXV (1967), 123–38.
Womack, J. P., D. T. Jones, and D. Roos, *The Machine That Changed the World: The Story of Lean Production* (New York, NY: Harper Perennial, 1991).