Technological Progress and Organizational Change: Appearance and Disappearance of the Hierarchy[1]

Micael Castanheira
ECARES, CEPR, and FNRS[2]

Mikko Leppämäki
Aalto University[3]

Date: April 3, 2023

Abstract: The paper addresses the optimal structure of a decision-making organization and examines how it is affected by technological progress. We show that technical evolutions that are not different in nature can have strikingly different effects on the organization structures. If the agents' productivity is initially low, as was the case in the 19th century, the productivity gains induce firms to become more hierarchical. By contrast, if the agents' productivity is initially high, as it is the case nowadays, the productivity gains induce firms to become flatter. The differences are only due to the different initial level of productivity. Our main result explains why technological progress induced organizations to adopt increasingly flattered ("delayered") structures only recently, even though it has been around forever. The model offers an explanation for the appearance and disappearance of the hierarchy. Interestingly, our model also predicts that the decision-maker's (CEO's) span of control increases in the latter phase of technological progress. That is, even if firms have become recently more delayered, the top management team (the layer just below a CEO) who directly communicates with a CEO has in turn increased. These theoretical findings coincide with the empirical evidence provided by Rajan and Wulf (2006) and Wulf (2012).

Keywords: hierarchy, theory of firm, organizational change, technological progress.
JEL Classification: D21, D73, L22

[1]We would like to thank Jacques Crémer for helpful discussion and Matti Keloharju, Deniz Okat and Frans Saxen for useful comments.

[2]ECARES, ULB CP 114, 50 Av. Roosevelt, 1050 Brussels, Belgium. E-mail: mcasta@ulb.ac.be.

[3]Aalto University, School of Business, Graduate School of Finance (GSF), P.O. Box 21220, FI-00076 Aalto, Finland. e-mail: mikko.leppamaki@aalto.fi

1 Introduction

Over the last 20-30 years, we have wittnessed a shift away from highly hierarchized organizations towards flatter structures. In practise, what we have oberved is called "Delayering": some layers of organizations has been removed; the top decision-maker (the CEO for instance) must get in closer contact with his or her top management team. In this paper we examine how delayering, CEO's span of control and technological progress interact within a theoretical model of decision-making organization.

Rajan and Wulf's (2006) analysis, which is based on a sample of 300 major US firms, identifies a tidal wave of organizational restructuring over the last 20 years. On average, CEOs used to only supervise 4.5 managers in 1986. In 1999, this number had increased to about 7: a 45% increase. One reason could be that the CEO's span of control had to increase because firm size increased. Instead, the average size of the firm had actually fallen from an average of 86000 to 70000 employees over the same period. Hierarchical structures have also become simpler: the typical US firm counts fewer layers of management. An open puzzle in that regard is that the number of tasks performed by the upper layers of management does not appear to have increased, in spite of the smaller number of layers below them.

Bresnahan, Brynjolfsson and Hitt (2002 – BBH henceforth) and Caroli and Van Reenen (2001) document that organizational restructuring might actually be a by-product of technical evolutions, and that they induce a concomitant shift toward higher-skilled labor force. According to BBH, the present technical evolutions, namely increased reliance on IT (information technology) services, is fundamentally different from the technical changes that occurred in the 19th century. Nineteenth century technical evolutions seemed to complement unskilled labor, whereas IT services require high-skilled labor as a complementary input. Therefore the current shift away from low-skilled workers, and the resulting upsurge in the wage gap between skilled and unskilled workers (or increased unemployment in Europe, see Caroli and Van Reenen, 2001). Whether or not there is a direct link between IT and delayering remains unclear.

This paper addresses the problem of organizational structure from a theoretical standpoint, and sheds new light on these evolutions. We show that, even technical evolutions that are not different in nature can have strikingly different effects on the organization. If the agents' productivity is initially low, as was the case in the 19th century, then productivity gains induce the firm to become steeper and more hierarchized. That used to be the best way to adapt to market conditions at the time. By contrast, if the agents' productivity is

initially high, as it is the case nowadays, then productivity gains induce the firm to become flatter. That is the best way to adapt to market conditions nowadays. These differences are only due to the different initial level of productivity. In a sense, this result explains why technological progress induced organizations to adopt increasingly decentralized and flatter structures only recently, even though it has been around forever.

Taking this broader perspective allows us to address another puzzle that has never been addressed by the literature: if it now appears optimal to implement delayering, why was it *not* optimal in the past? Expressed differently, there is little questioning about why organizations initially reached a state of intense hierarchization and low span of control for the CEO. The literature seems to take for granted that the developments of IT necessarily call for a flattening of the firm. Why is the organization's best response so different at different moments in time? This is the main contribution of our approach: we look at the evolution of organizations with a longer-term perspective. We not only explain the most recent and documented organizational changes. We also provide a encompassing explanation for why hierarchization takes place in the early stages of a technology and flattening occurs in a second phase.

We model the problem of an organization as the need to perform a given set of tasks. Both the CEO and the agents need time to analyze and solve the problems they have at hand. Whenever a problem is solved, the organization moves on to another problem, and so on until all problems are solved. Borrowing from the information processing literature (see e.g. Radner and Van Zandt 1992, Radner 1993, Bolton and Dewatripont 1994), one can rationalize the time it takes to make a decision by some bounded rationality argument. Bounded rationality then allows to explain why a given person may become overloaded if she has to make all decisions by herself. For that reason, it may become optimal to delegate some tasks to subordinates. This need to delegate tasks, while coordinating the agents, explains how and why a given organization should be set up, and which is its optimal structure.

Main results

In this paper, we take a much simplified approach. Our goal is not so much to rationalize why the optimal organization should or not be a hierarchy, nor to find the optimal communication structure inside the firm. Instead, we develop a model in which, by construction, a hierarchy is optimal. Still, the model will allow us to make precise predictions about the optimal structure of this hierarchy: how many layers of management should it count? What is the optimal span of control of the CEO? How should it adapt following a change in technology?

Since our main focus is on technical evolutions, we simplify as much as possible the representation of the way in which decisions are made. Each agent is able to solve a given fraction of the problems that he is facing. To repeat, delegating tasks to that agent means that he will make some decisions, which do not have to be cross-checked by his superiors at later stages. Instead, only the problems that remain unsolved have to be relayed to other layers of management, in a fashion similar to Garicano (2000). In contrast with Garicano (2000), however, we leave exogenous the number of problems that can or cannot be solved by the agent, and we impose symmetry in the capacity of the agents' ability to solve problems.

This simplifies tremendously our analysis and allows us to focus on that specific aspect of decision-making capacity that has been overlooked until now. Namely, how will the optimal organization evolve if, thanks to some exogenous evolution, an agent becomes able to solve more problems? We identify two opposite forces. First, what we call a "*job creation*" effect: since the agents are becoming more productive, it becomes worthwhile hiring more of them. For that reason, the organization should expand. We show that this expansion typically materializes into an increased hierarchization, together with a reduction in the CEO's span of control. Second, there is a "*job destruction*" effect that works in the opposite direction. Since a given agent is becoming more productive, he leaves fewer problems unsolved. This reduces the amount of work to be performed by the upper layers in the organization, and therefore reduces the need to hire workers in these layers. Typically, this force induces the organization to become flatter, and causes the CEO's span of control to increase. Overall, CEO's span of control is thus U-shaped in productivity.

The open question is when the former –job creation– or the latter –job destruction– effect comes to dominate. Our results demonstrate that the main driver of change is the initial level of productivity of the agents. If the agents are only able to make few decisions initially, then an increase in their productivity is essentially job creating. Conversely, if the agents are initially able to make a relatively large number of decisions, then an increase in their productivity is essentially destroying jobs. We thus find a hump-shaped relationship between the total number of workers (and layers) and technological progress (captured by the increase in labour-embodied productivity).

It is useful to note that at the linkages between technological progress and employment are not always clear at the macroeconomic level either. In the endogenous growth literature, technological progress generates an expansion in economic activity – albeit for technological obsolescence reasons (see e.g. Aghion and Howitt 1997). In some cases, technological progress can also generate a "skill bias" and act against some fringes of the population (see e.g.

Acemoglu 1998). In this paper instead, we assume perfectly homogenous agents whose task consists of easing decision-making by the firm's management. Though very simple, this framework allows us to explain why technological progress has opposite effects on the profits of the organization and on employment prospects, and why organizations should become "flatter" when economic conditions worsen. Nevertheless, the evidence of BBH shows that, in present times, technical evolutions require the use of human capital to bear a productive fruit. Interpreted in that way, our results thus show that the firm values positively human capital (that generates labor-embodied productivity gains) and negatively labor (since fewer agents must be hired). In any case, we demonstrate that organizational changes are highly complementary to technical progress, even in the absence of skill-biased technological change.

Related literature

We already referred to a few empirical studies that provide crucial insights into the interactions between technical progress, span of control, and organizational structure. As we have seen, most of their findings are widely consistent with the predictions of our model. Moreover, our results span further, by explaining why the initial situation could have emerged, and why the firms' adaptation trend has reversed in the meantime.

There are other, complementary, analyses that look at the linkages between technologies and organizational structure. Crémer, Garicano and Prat (2005) propose a model of codes in organization, in which it is the ability to transmit accurate messages that determines the optimal shape of the organization. A declining price of technology allows the firm to ease communication, adopt a different coding system, and therefore a different organizational structure. Garicano and Rossi-Hanberg (2003) provide another insightful analysis of organizations in a knowledge economy. They examine how information technology affects wages and organizations, and explain a series of commonly observed phenomena such as sorting by ability or the relationship between rank and cognitive ability. Prat (1997) proposes another model of the organization, in which technology is entirely dependent on human capital. He shows that, depending on the wage structure, the organization will have more or fewer layers, with more qualified workers 'on top'. Qian (1994) is another important and interesting analysis of optimal hierarchies, but does not either address the question of how technological progress would shape the size and optimal structure of hierarchies.

The remainder of the paper is structured as follows. Section 2 lays out the model of the decision-making organization. In Section 3, we examine the organization's optimal response to exogenous changes in the cost of labor, relative to the benefit of performing more efficiently. In our setup, efficiency in decision-making essentially amounts to being able to

reach a decision rapidly. We show that when the urgency of decision-making becomes more important for the success of the organization, the optimal response is to hire more workers and structure the organization in a more hierarchized manner. By contrast, when it becomes comparatively more important to reduce the wage bill, the organization should hire fewer workers and have a flatter organization. Note that these variations come on top of the organization's best response to technical evolutions: if wage costs become an issue, the organization becomes flatter than the structure that would maximize its ability to reach a decision rapidly. Section 4 addresses our main question. We examine the effects of technological progress on organizational structures and the top managements' span of control. In Section 5, we consider an alternative mode of organization: just-in-time. There, we show the high complementarity between technology and organizational structure, with a result that helps explain the productivity paradox. Finally, Section 6 concludes.

2 The Model

We model the problem of a CEO, or *decision-maker* (she), who must decide whether to implement a project. To this end, a number of decisions have to be made. We represent the set of initial options by a continuum with mass M. All elements within this set must be processed, and a set of decisions with mass M must be reached. For simplicity, we can assume a simple 'yes/no' type of decision: only a few number of elements in this set are worth pursuing.[1]

Decision-maker vs. Agents. The decision-maker is the person who eventually implements projects. Profitable implementation depends on her own skills and, therefore, she is the only person who can materialize the value of a project. She also has a "comparative advantage" at pinpointing which projects must be implemented. However, project evaluation is a time-consuming task. This means that, if she worked alone, the decision-maker would need an amount of time M to make all decisions: she would have to assess the returns of each project, and decide which ones to implement. Hence, in spite of her comparative advantage, she may benefit from delegating some decision powers to external agents, to speed up decision-making.

We normalize to 1 the amount of time needed to process a mass 1 of projects, for any person in the organization (decision-maker or agent). The decision-maker thus needs M_0

[1] By assumption, implementing a "good" project generates an arbitrarily large surplus, whereas implementing a "bad" project generates arbitrarily large losses. The purpose of this assumption is only to ensure that it is optimal to assess the return of all elements in the set, before deciding which one(s) to implement.

units of time to identify which projects she to implement, if she has a mass M_0 left to screen by herself.

Although agents take the same amount of time to evaluate a project,[2] they are never able to perfectly identify whether the decision-maker could turn a given project into a profitable venture. When an agent evaluates a project, he obtains a noisy signal s about its value.

For simplicity, we assume that this signal has a binary structure: $s \in \{s_+, s_-\}$, where s_+ conveys the information that the project is potentially valuable, and s_- conveys the information that the project has negative value. The probability to receive each of these signals depends on the actual value of the project under scrutiny: if the project is actually worth implementing, then $s = s_+$ with probability 1. If the project has negative value added, then $s = s_+$ with probability f, and $s = s_-$ with probability $(1 - f)$. In other words, after receiving signal s_-, the agent can update his beliefs and infer that the project has negative value added with probability one.[3] It must be discarded, and should not be transmitted higher up in the hierarchy.[4] Thus, $(1 - f)$ can be interpreted as the *productivity of the agent*, and f as the *failure rate* of agents in the organization.

By the law of large numbers, the fraction of projects that an agent can filter out is $(1 - f)$. This is the number of decisions he can make. Throughout the paper, we treat this number as an exogenous parameter. Conversely, an agent leaves a fraction f still to be screened by his superior. Importantly, these fractions are assumed independent of the hierarchical position of the agent, and of the number of agents in the organization.[5]

[2] This assumption is akin to the one in Bolton and Dewatripont (1994), where the time needed to process a given amount of information is directly proportional to the mass of information.

[3] That is, agents can only make "type-I" errors. See Sah and Stiglitz (1986) for a model in which agents can also make "type-II" errors.

[4] We assume away moral hazard concerns of the types considered by Aghion and Tirole (1997): by assumption, agents have no incentive to favor projects that are not profitable to the decision-maker, and do benefit from revealing that a project is potentially profitable.

[5] This assumption borrows from Radner (1993), where agents must add numbers. The time taken to add up n numbers two by two is assumed constant and equal to n. In our setup, f could be re-interpreted as the inverse of the number of operations performed by an agent, per unit of time.

Hierarchy. To coordinate agents, the decision-maker can create a hierarchical organization, which is structured into *layers*, denoted by $l = 0, ..., L$. In such a hierarchy, the initial set (M) is first processed in layer L, then in layer $L-1$, $L-2$, and so on, until the top of the hierarchy (the decision-maker, in layer 0) is reached.[6] Since each agent's decisions remove a fraction $(1-f)$ from the set he works on, and since he transmits the remaining fraction f to his superior, the mass of items reaching layer l is equal to $M \times f^{L-l}$. By increasing the number of layers, the decision-maker thus reduces the mass of information that she must process by herself, to $M \times f^L$. On the other hand, she also delays the moment at which she can initiate her own work.

In addition to the number of layers in the hierarchy, the decision-maker controls the number of agents, n_l, who work in each layer l of the hierarchy. Since the failure rate of agents f is independent of how work is divided among the agents, the decision-maker can reduce the amount of time they spend evaluating projects, by increasing the number of agents in the layer. This delay reduction captures the benefits of increased division of labor. On the other hand, increasing the number of agents increases the time needed to coordinate their tasks; we know that big organizations can become quite difficult to manage, partly because of coordination and communication costs (Radner 1993, Bolton and Dewatripont 1994). Like Radner (1993), we introduce such costs in a reduced form, and assume that each agent slows down the organization by some fixed delay ϕ.[7] We call ϕ the *per agent coordination cost*. Taking these two effects into account, layer l needs:

$$d_l = \frac{M_l}{n_l} + \phi n_l \tag{1}$$

units of time to process a mass of information $M_l \equiv M \times f^{L-l}$. The first term in (1) represents the time needed by the agents to process M_l (each processes a fraction $1/n_l$ of it); the second term represents total coordination costs in the layer.

Total delay. Typically, the information processing literature assumes that a superior can only start processing the information when his or her subordinates have completed their task. In Bolton and Dewatripont (1994), for instance, a superior must receive the summary written by his subordinates before being able to read and process them. In our model, we call this assumption the *sequential technology*: we assume that each layer works in sequence.

[6] We do not allow for "skip-level reporting" (See Radner 1993, Bolton and Dewatripont 1994).

[7] Radner (1993) interprets this parameter as the *unit cost of a processor*. We simply express this cost in the same units as delay. Note also that communication costs *à la* Dewatripont-Bolton could be considered, if we were assuming that the cost of adding one agent in layer l were some function $\phi(n_l, n_{l-1}, n_{l+1})$, which they derive from the fundamentals of their model. Introducing such functional forms substantially complexifies our results, in a way that is orthogonal to the focus of this paper.

This sequential technology amounts to imposing that layer l only starts working when layer $l+1$ has finished the processing of all M_{l+1} items. Accordingly, the total delay needed by the organization, D, is the *sum* of the delays needed by each layer:

$$D\left(L,\{n_l\}\right)=\sum_{l=1}^{L}\left(M\frac{f^{L-l}}{n_l}+\phi n_l\right)+M\,f^L. \tag{2}$$

The first term in (2) is the sum of the delays imposed by each layer of agents. The last term represents the time needed by the decision-maker to decide about which project(s) will actually be implemented.

We contrast this sequential technology with *just-in-time* technology in Section 6 in Appendix. Under the latter technology, the decision-maker has the possibility to make all layers work at the same time. Many organizations have been restructured to grasp the benefits of such just-in-time processing in the last decades, and comparing the two technologies will allow us to analyze some of the implications of this organizational switch.

Objective function and costs. The objective of the decision-maker is to reach her decision at minimum cost. Total costs stem from the time needed to reach a decision, and from the agents' wage costs. For the sake of tractability, we assume that the marginal cost of delay is constant and equal to r.[8] Similarly, wage costs are proportional to the amount of time worked by the agents: an agent who processes a mass m_i will be paid a wage $w\,m_i$. The total wage bill W thus amounts to:

$$W\left(L,\{n_l\}\right)=w\sum_{l=1}^{L}n_l\times\frac{M}{n_l}\,f^{L-l}=wM\,\frac{1-f^L}{1-f}. \tag{3}$$

In light of (2) and (3), the objective function of the decision-maker is thus to minimize the total cost function:

$$\min_{L,\{n_l\}}\ TC\left(L,\{n_l\}\right)=r\,D\left(L,\{n_l\}\right)+W\left(L,\{n_l\}\right). \tag{4}$$

3 Optimal Shape of the Organization

The problem faced by the decision-maker can thus be summarized with four exogenous parameters and $L+1$ control variables. The exogenous parameters are the efficiency of delegation,

[8] The assumption of a constant marginal cost of delay may seem unrealistic, but it is sufficient for the purpose of our analysis. In a more general setting, the marginal cost of delay could be represented by some function $r\left(D\right)$, and an increase (respectively decrease) in the marginal cost of delay be captured by setting $\tilde{r}\left(D\right)>r\left(D\right)$ (resp. $\tilde{r}\left(\cdot\right)<r\left(\cdot\right)$), $\forall D$.

proxied by f, the per-agent coordination cost ϕ, and the two marginal cost variables, w and r. The controls are the number of layers in the hierarchy L, and the number of agents in each of these layers: n_l, with $l = 1, ..., L$.

Working recursively, we first derive the optimal number of agents in each layer, holding fixed the number of layers. Next, we derive the optimal number of layers, and thereby obtain the optimal shape of the organization:

Lemma 1 *Given an exogenous number of layers L, the optimal number of agents in layer l is given by:*

$$n_l^* = \sqrt{f^{L-l}\frac{M}{\phi}},$$

and is thus independent of the marginal cost variables w and r.

Proof. *Neglecting integer constraints, by* $(2) - (4)$, *the first order condition is:*

$$\frac{\partial TC\left(L, \{n_l\}\right)}{\partial n_l} = r\left(-M\frac{f^{L-l}}{n_l^2} + \phi\right) = 0, \tag{5}$$

which immediately yields n_l^*. Further differencing (5) *wrt n_l shows that the second order condition is satisfied as well.* ■

The result that the optimal number of agents in a layer is independent of wages may seem surprising at first glance. Yet, the intuition is straightforward. By (3), one observes that, for a given number of layers, the wage cost is independent of n_l: since layer l must process a fixed mass of information $M\, f^{L-l}$, the total amount of time worked by the agents in that layer is also $M\, f^{L-l}$, by definition. Changing the number of agents in the layer only affects how this workload is shared among the agents in l: either a small number of agents work for a long time, or more agents work for a smaller amount of time. Hence, the number of agents only affects the delay needed by the layer to perform its task, leaving wage costs unchanged. This is why only the ratio between the mass of information and coordination costs matter.

Using Lemma 1, one can simplify the expression of the total cost function into:

$$\begin{aligned} TC\left(L, \{n_l^*\}\right) &= 2r\sum_{l=1}^{L}\left(\sqrt{\phi M\ f^{L-l}}\right) + rM\, f^L + wM\,\frac{1-f^L}{1-f} \\ &= 2r\sqrt{\phi M}\,\frac{1-\sqrt{f^L}}{1-\sqrt{f}} + M\,\left(r\, f^L + w\frac{1-f^L}{1-f}\right). \end{aligned} \tag{6}$$

Having optimized for the number of agents in each layer, the only remaining control variable is the number of layers in the organization. (6) shows which trade-off emerges regarding

the number of layers in the organization. On the one hand, increasing the number of layers reduces the time worked by the decision-maker: $M_0 = M f^L$ is strictly decreasing in L. However, this is obtained at the expense of longer delays needed by the agents, and of an increased wage bill (the first and the last terms in (6) are strictly increasing in L).

Below, we use the superscript '**' to denote the optimal level of a variable, and the superscript '*' to denote the optimal level of a variable when the number of layers is restricted to be an integer. Exploiting Lemma 1 and the trade-off highlighted by (6), we are now in a position to determine the shape of the optimal hierarchy:

Proposition 1 *Optimal hierarchies are necessarily pyramidal* (*i.e.* $n_{l-1} \leq n_l$), *and the* span of control *of each agent in a layer* $l < L$ *is given by* $1/\sqrt{f}$. *The optimal number of layers in the organization is given by:*

$$L^* = \max\left\{\left\lceil 2\frac{\log\left(2\sqrt{\phi/M}\right) - \log[Z]}{\log[f]}\right\rceil, 0\right\}, \tag{7}$$

$$\text{where } Z = (1-f) - w/r \tag{8}$$

which is increasing in M/ϕ, *and decreasing in* w/r. *Abstracting from integer constraints, the total size of this organization is:*

$$N^{**} \equiv \sum_{l=1}^{L^{**}} n_l^* = \frac{Z^2 - 4\phi/M}{(1-f)\ Z^2\ \sqrt{\phi/M}}, \tag{9}$$

which is also increasing in M/ϕ *and decreasing in* w/r.

Proof. *That optimal hierarchies are pyramidal follows immediately from Lemma 1:*

$$\frac{n_{l-1}^*}{n_l^*} = \frac{\sqrt{f^{L-l+1}\frac{M}{\phi}}}{\sqrt{f^{L-l}\frac{M}{\phi}}} = \sqrt{f}.$$

*The span of control of each agent (*i.e. *the number of subordinates he has) in a layer* $l < L$ *is thus equal to* $1/\sqrt{f}$.

To compute the optimal number of layers, we proceed in two steps: first, neglecting integer constraints, we use (6) *to compute how total costs would vary if the total number of layers was increased by 1, and derive the value* L^{**} *that sets this difference equal to zero:*

$$TC(L+1, \{n_l^*\}) - TC(L, \{n_l^*\}) = 2r\sqrt{\phi M f^{L^{**}}} - M(r(1-f) - w) f^{L^{**}} = 0.$$

From this difference, it follows that total costs are monotonically increasing in L if $r(1-f) < w$. In that case, $L^{**} = 0$. Conversely, when this difference is negative for $L^{**} \to 0$, we get:

$$L^{**} = 2\frac{\log\left(2\sqrt{\phi/M}\right) - \log\left[(1-f) - w/r\right]}{\log\left[f\right]} \tag{10}$$

Then, reintroducing integer constraints yields (7)*, where $\lceil \cdot \rceil$ denotes rounding up to the nearest integer. Moreover, it is straightforward to check that:*

$$\frac{dL^{**}}{d\phi} < 0; \ \frac{dL^{**}}{dM} > 0; \ \textit{and} \ \frac{dL^{**}}{dw/r} < 0.$$

To derive the optimal size of the organization, we sum the optimal number of agents in each layer as given by Lemma 1, for the optimal number of layers L^{**}. Noting that: $f^{\frac{K}{\log(f)}-l} = \exp(K)\ f^{-l}$, straightforward algebra yields (9)*.* ■

Thus, in our framework, any optimal hierarchy is pyramidal (see also Bolton and Dewatripont 1994, Garicano 2000, Prat 1997, and Qian 1994).[9] What is more, it has the properties of a *preprocessing/tree network* (Radner 1993, pp1123-1124); *i.e.* all agents in a given layer have the same number of direct subordinates (which is inversely related to the failure rate: $n_{l-1}/n_l = 1/\sqrt{f}$) as well as indirect subordinates.

Proposition 1 also identifies how the shape of the organization responds to the environment external to the organization. If we interpret r as reflecting the strategy of the organization, it reveals the manager's valuation of faster processing, given the actions of competing organizations. In a highly competitive environment, it may be crucial for the organization to quickly react to new elements of information and/or to be the first in a patent race. In that case, the manager will infer that it is comparatively more important to reduce delays than cut wage costs (w/r is small). In that case, her organization will be more hierarchized, employ more people, and perform faster (Proposition 1 and Corollary 1). Conversely, if price competition dominates, speed may matter less than cost reductions. In such environments, the firm will behave as if the shadow cost of employment were high, and the marginal cost of delay low (w/r is high). In that case, the decision-maker must opt for a flatter structure, at the expense of slower processing:

Corollary 1 *Integer constraints notwithstanding, the optimal processing delay is given by:*

$$D^{**} = f^{L^{**}}M + 2\frac{1-\sqrt{f^{L^{**}}}}{1-\sqrt{f}}\sqrt{M\phi} = \frac{4\phi}{Z^2} + 2\frac{\sqrt{M\phi} - \frac{2\phi}{(1-f-w/r)}}{1-\sqrt{f}}, \tag{11}$$

[9] This also means that, with the simplifications we introduced, we cannot explain other organizational structures, such as unbalanced networks (see Radner 1993) or conveyor belt organizations (see Bolton and Dewatripont 1994).

where $Z = (1 - f) - w/r$. *Delay is thus increasing in* M, ϕ, f, *and* w/r.

The predictions of Proposition 1 are widely consistent with the evidence provided by Rajan and Wulf (2006): they find a significantly negative correlation between:

a) the decision-maker's span of control (n_1 in our model; see next section for more detail) and the number of layers in the organization ('depth'), and

b) the number of layers and the wage of the agents.

While we defer discussion about the decision-maker's span of control to the next section, Proposition 1 provides a direct rationale for the second correlation: when external competition induces the decision-maker to contain wage costs, she should not decrease the number of agents in each layer, but instead reduce the number of layers (intuitively, the former strategy causes the number of agents in layer 1 to increase, whereas the latter strategy would have caused it to decrease, which accords to their evidence).

Using another possible interpretation of the parameters w and r, one may nevertheless question the congruence of our results with these correlations. Indeed, shouldn't a straightforward interpretation be that an exogenous increase in wages must cause the number of layers to decrease, as illustrated in Figure 1?

To address this question, it is important to note that the specification of Rajan and Wulf's regressions either includes year dummies or a time trend, depending on the specification. Therefore, the partial correlation between the number of layers and the divisional managers' wages (the wage of the agents in our setup) essentially establishes a cross sectional correlation. By contrast, the effect of time on the number of layers is negative and significant: firms became flatter over time.

Since the wage of a manager has had a positive trend over the 1990s, one must draw the conclusion that, over time, there is indeed a negative correlation between the aggregate wage level and the number of layers in the organization (in addition, note that Rajan and Wulf reveal a decline in organizational size over the same period, as the right-hand graph in Figure 1 suggests). However, once this aggregate correlation has been controlled for, there remains a negative correlation between wages and organizational 'depth'. Our results suggest that this reveals the decision-theoretic nature of the w/r ratio: it reveals the strategic choices of the firm; its relative stress on performance (reduced delay) versus cost containment (wages cuts).

4 Technological Progress/Technical Change

Beyond the organizational changes that directly relate to the relative cost of wages, increasingly wide evidence reveals that "delayering" goes along with increased empowerment of the managers in lower layers. Furthermore, the returns on investing in Information Technologies essentially seem to materialize when such organizational changes are implemented: organizational change is complementary to technological evolutions (Brynjolfsson and Hitt 2003, Caroli and Van Reenen 2001, Rajan and Wulf 2003). This means that, either by means of the organizational structure, by means of improved technologies (e.g. the implementation of Information Technologies), by means of hiring more able agents, or by a combination of these, firms increase their productivity by increasing the number of decisions made by lower-level managers. The productivity of the agents ($1 - f$ in our setup) has thus been increased progressively, and this empowerment of the labor force in turn prods a shift in the composition of the employed labour force in favor of more highly-skilled workers (see e.g. Acemoglu 1998, Autor *et al.* 1998, Caroli and Van Reenen 2001, or BBH for an analysis of these changes on the skill-bias). Hence, a conclusion that emerges is that the agent's productivity level is a prominent factor to explain the shape of the firm. To a large extent, technical evolutions explain organizational change as well. In particular, recent improvements in the employees' productivity are the main cause of the flattening of the firm.

However, productivity increases are not a recent phenomenon, while the flattening of the firm is. Why is it then that, in the more distant past, the effects of productivity improvements were exactly the opposite? Think indeed about the industrial revolution: it is thanks to similar productivity improvements that the industrial revolution transformed small workshops into strongly hierarchized corporations. These organizations, composed of a high number of layers and agents, dominated the scene until recently. Now, a reversal occurred and "flattening" has become the prevalent strategy.

An appropriate model of the organization must be able to explain both evolutions: why did we initially witness a trend of increasing hierarchization, now followed by an opposite trend of "delayering"? Why was it initially optimal to centralize core tasks in a top management layer, while it now proves more efficient to reduce the number of layers? Why is it now optimal to decentralize tasks throughout remaining layers, and centralize remaining tasks directly in the hands of the corporation's CEO? Why is it optimal for some firms to maintain a strongly hierarchized structure, while other firms opted for intensive delayering?

4.1 Capturing Technical Progress

To address these issues, we have to identify how technological progress can be captured in our model. We can measure technological progress along several dimensions. First, it may simply become quicker to perform any task. In that case, equation (2) that describes how delay is computed should be adapted: delay should be divided by a factor measuring processing speed. However, so must be equation (3) that characterizes total wage costs. Per se, higher processing speed can thus not affect the shape of the organization; it reduces delays and wage costs in proportion, and leaves unchanged the optimal shape of the hierarchy. This is the typical effect of disembodied technical change: productivity increases but does not affect the relative demand for skills nor labor.

A second aspect of technological progress that our model can capture is the easiness to coordinate large teams of workers. This is measured by the parameter ϕ: the design of modern factories in the nineteenth century, like the spread of information technology in the late twentieth century, were partly meant to lower monitoring and coordination costs. According to Proposition 1, this induces an increase in the optimal number of layers, as well as in layer size (see Lemma 1 and Proposition 1). Corollary 1 also shows that containing ϕ allows to reduce processing delays. A reduction in ϕ is thus consistent with the developments that occurred during the industrial revolution (increased division of labour, and hierarchization of the organization), but not with more recent trends.[10]

Third, the increased capacity to process information could be captured through the mass of information that is processed. In the past, corporations tended to develop all aspects of their business "in house". More recently, this trend has reversed. Corporations now tend to focus on their "core business", and to outsource side activities (see *e.g.* Domberger 1998, Grossman and Helpman 2002, or Castanheira and Leppämäki 2003). If these two evolutions are represented respectively by an increase and a decrease in M, the model would generate a dynamics of the organizational structure that is consistent with stylized facts. Consider for instance the developments in the car industry (Womack, Jones, and Roos, 1991). Technological progress initially helped designing more sophisticated and complex cars (M increases). Then, the lean production revolution induced car producers to reduce the number of parts of their products (M decreases). Again, Proposition 1 shows that such an evolution must generate organizational changes that are consistent with observations. Such

[10] See Crémer, Garicano, and Prat (2005), for a model in which a drop in the cost of communicating messages among agents induces the organization do change shape, because the type of code used in the organization can be changed as well.

a result can be traced back, again, to the seminal paper by Radner (1993). However, this leaves unaddressed the question of why technical evolutions would have, first, incentivated the organizations to expand M in this way and, second, to try and reduce the scope of their activities.

In the remainder of this section, we aim to show that a unique factor, namely labor-embodied productivity, can be at the heart of these –apparently contradictory– evolutions. If labor-embodied productivity is initially low, technical progress induces the organization to expand and become more hierarchized. By contrast, if labor-embodied productivity is initially high, further technical progress induces the organization to become flatter. Surprisingly, this labor-embodied aspect of the efficiency of task delegation did not receive much attention in the literature.

4.2 Increasing Labor-Embodied Productivity

For the sake of concreteness, let us begin with the potential effects of Information Technologies (IT). Relying on IT should increase labor-embodied productivity, by allowing the agents to process information more easily. In this sense, one must draw a parallel between technological progress (the spread of IT) and the capacity of the agents to perform managerial and processing tasks: as state Brynjolfsson and Hitt (2000, p24), "[t]*he fundamental economic role of computers becomes clearer if one thinks about organizations and markets as information processors (Galbraith, 1977; Simon, 1976; Hayek, 1945). Most of our economic institutions and intuitions emerged in an era of relatively high communications cost and limited computational capability.* [...] [It] *is not surprising that the massive reduction in computing and communications costs has engendered a substantial restructuring of the economy,* [...] *some of the most interesting and productive developments* [being] *organizational innovations.*"

Their evidence shows that the delegation of decision tasks has widely increased along with the spread of IT in organizations. Agents lower in the hierarchy are given increased decision powers; they have the right (and, by implication, the capacity) to remove more items from the initial information set. Since they make more decisions, they transmit fewer items to their superiors. This, in our model, is captured by a value of f that is falling over time. Quite intuitively, IT opened the gates to such increased delegation of powers: the decision-maker can assign and delegate tasks more easily; she can spread more widely the details about how a project should be evaluated, and therefore reduce the number of mistakes made by the agents.

Our second proposition aims at identifying how such technological progress influences the shape of the organization. It highlights the effect of an increase in $(1-f)$ on the number of agents reporting directly to the decision-maker (her span of control), and on the overall shape of the hierarchy:

Proposition 2 *i) Integer constraints notwithstanding, the decision-maker's span of control is given by:*

$$n_1^{**} = \frac{2}{(1-f-w/r)\sqrt{f}}.$$

*ii) For w/r sufficiently low and M sufficiently large, an increase in the efficiency of task delegation $(1-f)$ generates **job creation** and **higher hierarchization** in the early phases of development (f initially close to 1), and **job destruction** as well as **decreasing hierarchization** in the later phases of development (f closer to 0). In the long-run, it generates "flat" hierarchies ($\lim_{f\to 0} L^* = 1$).*

Proof. By Lemma 1 and Proposition 1, the optimal number of agents in the first layer is given by:

$$\begin{aligned} n_1^* &= \sqrt{f^{L^{**}-1}\frac{M}{\phi}} = \left(f^{2\frac{\log\left(2\sqrt{\phi/M}\right)-\log[(1-f)-w/r]}{\log[f]}}\frac{M}{\phi f}\right)^{\frac{1}{2}} \\ &= \left(\frac{4\phi/M}{(1-f-w/r)^2\, f}\frac{M}{\phi f}\right)^{\frac{1}{2}}. \end{aligned}$$

Next, from (7), we see that $L^* = 0$ for $f \geq 1-w/r$ and $L^* = 1$ for $f \to 0$. Differentiating (10) with respect to f also shows that $\frac{dL^{**}}{df} < 0$ iff

$$(1-f-w/r)\left(\log\left[2\sqrt{\phi/M}\right] - \log\left[1-f-w/r\right]\right) > f\,\log\left[f\right],$$

which is satisfied for $f \to (1-w/r)$ and violated for $f \to 0$ (note that $\lim_{x\to 0} x\,\log[x] = 0$).

Applying the same reasoning to (9) proves the proposition. ■

Proposition 2 shows that the decision-maker's span of control is independent of the mass of decisions to be made, M, as well as of per-agent coordination costs, ϕ. As seen in Proposition 1, these two parameters do affect the optimal shape of the organization, but the latter adapts in such a way that the number of agents reporting directly to the decision-maker remains independent of these parameters. In a sense, this confirms our initial intuition that there is something more than an exogenous change of M over time that explains the organizational evolutions that took place over time.

By contrast, the productivity of delegation $(1-f)$, as well as relative wage costs, w/r, do affect the decision-maker's span of control. The first part of Proposition 2 confirms the intuition we developed in the previous section, that the decision-maker's span of control should be increasing in the relative wage cost. In light of the discussion of Proposition 2, one must understand that this is a side-effect of the firm's 'cost-cutting delayering' – it is meant to contain wage costs at the expense of longer delays. Compared to the fastest organization (the one that would obtain when $w/r \to 0$), the number of layers is reduced, as is total employment, but the top management layer becomes increasingly large. This result is perfectly in line with the evidence provided by Rajan and Wulf (2006).

Interestingly, Proposition 2 also shows that the decision-maker's span of control is not monotonic in the efficiency of task delegation, $(1-f)$: if we track the evolution of the decision-maker's span of control when $1-f$ increases from 0 to 1, we see that n_1^{**} is initially *de*creasing in $1-f$ and reaches a minimum in $f = 1/3$. For larger values of $1-f$, instead, her optimum span of control must be increasing in the agents' productivity. Figure 2 below illustrates this pattern.

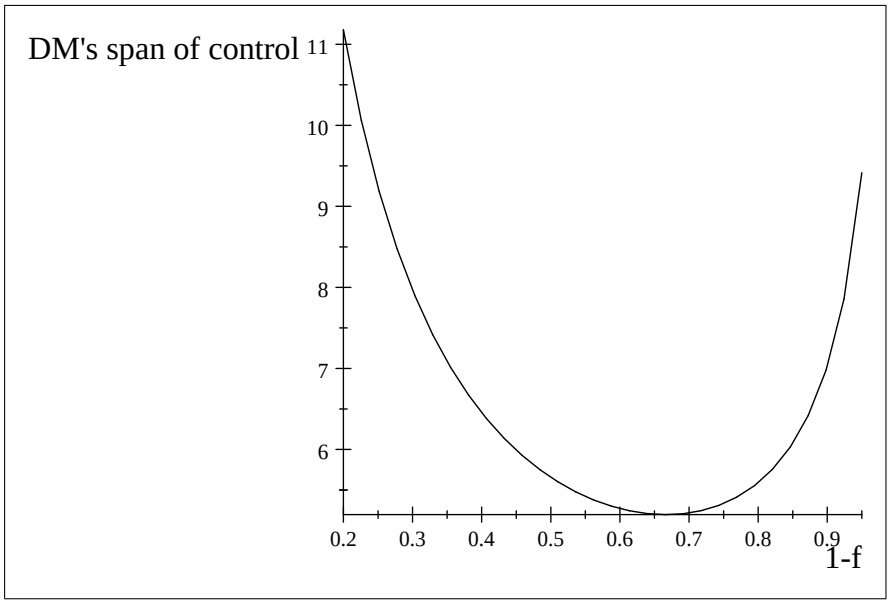

Figure 2: Effect of the efficiency of task delegation on the decision-maker's span of control.

To understand this U-shaped relationship, it is useful to relate this result to the second part of Proposition 2. When productivity of delegation is very low (f close to 1), the decision-maker avoids delegating *any* task: agents would take time to perform their tasks, and the result of their effort would be of little use anyway. Therefore, the optimal organization is composed of the decision-maker alone ($L^* = 0$). This case can be interpreted as the

self-employed craftsman's "organization". As the productivity of delegation increases, our craftsman gains interest in taking advantage of division of labour and increased task specialization. Now, she creates a one-layer organization ($L^* = 1$), in which she directly supervises all employees (on Figure 2, this is the left-most part of the graph). These (potentially numerous) employees work in parallel and, by revealed preferences, speed up processing and reduce total costs. Further improvements in productivity generate more dramatic organizational changes. When the agents' productivity further increases, a complete hierarchy gets created: the decision-maker reduces the number of agents that are under her direct supervision (see Figure 2), but keeps increasing total employment (see the right-hand pane of Figure 3, below); she delegates to a "top management team" the authority of organizing and supervising the work of subordinate agents. Increased productivity has a *job creation effect*.

Along these changes, note that each agent in a layer $l < L$ supervises an increasingly large number of direct subordinates, and makes an increasingly large number of decisions. By implication, if the optimal number of layers is increasing, the total size of the organization (the total number of employees) must be increasing as well, and so is the overall level of delegation (M_0/M decreases: the decision-maker performs fewer tasks by herself). Moreover, the ratio $\sum_l n_l/n_1$ keeps increasing in $(1-f)$: top managers have an increasingly large number of direct and indirect subordinates, since there are more layers under their authority.

Yet, this process is far from being monotonic: for intermediate values of f (around $(1-f) = 1/3$ in the simulations used for Figure 3), the optimal number of layers, as well as the number of employees, starts falling. The rationale for this result is quite simple. Since lower layers in the organization are gaining efficiency, they perform a larger fraction of tasks by themselves: decision-making has become more *decentralized*.

Hence, fewer tasks remain to be performed at higher levels of the hierarchy: fewer people, and eventually fewer layers, are needed in the organization. This is a *job destruction* effect of increased productivity. Summing up, depending on the initial productivity level, technological progress can either be a blessing or a curse for the workers of this organization. If productivity is initially very low, technological progress allows the creation of new job opportunities in the organization. However, if the agents' productivity is already high, then technological progress becomes detrimental to employment: the same tasks can be achieved with fewer agents, and in flatter hierarchies (see Figure 3 below). In the limit, when $1-f$ approaches 1, the optimal number of layers converges to 1: the organization becomes perfectly flat, decisions are fully decentralized. The number of agents is then determined by Lemma 1, for $L^* = 1$.

Still, delay and total costs are monotonically decreasing in $(1-f)$, which means that the decision-maker always benefits from technological progress accompanied by appropriate organizational adaptation. Noticeably, the decision-maker's "demand" for technological improvements will be more intense in organizations where many tasks are already delegated, and in which total employment will concomitantly be reduced (when productivity is initially sufficiently high).[11]

Note that this result is quite robust. It does not hinge on the way in which we modelled the organization; even though our modelling approach allowed us to highlight this result. If one instead considers a model à-la-Radner, where the organization has to process K information items, then the optimal organization would be composed of only one agent if this agent is able to process all items in one cycle. Similarly, in a model à-la-Garicano, if it becomes costless to acquire skills that allow an agent to treat all type of tasks that have to be executed in the organization, the optimal organization would become flat as well. The essence of our result is only that we managed to isolate a fundamental effect of increasing the productivity of an agent: the more tasks he can perform by himself, the fewer tasks remain to be performed in other layers in the hierarchy; the job destruction aspect of higher productivity. In a Radner-type of model, because the number of items to be processed is countable, the optimal organization shrinks to one individual for a purely mechanical reason. In other setups, such as the one used in Garicano (2000) or in this paper, this mechanical effect is set aside. Yet, the dual effects of increased productivity remain present: on the one hand, it becomes more valuable to delegate tasks, since the agents perform better; on the other hand, when their productivity is sufficiently large, *they* should not delegate tasks to other agents. In our setup, since the organizational design is chosen by the decision-maker, this is exactly what happens: the organization becomes leaner and flatter. If some of the organizational design were left in the hands of the agents themselves, additional effects would have to be considered: delegation reduces the firm's performance, but allows the agent to perform fewer tasks by himself as well.

Remark 1 *These results also provide a hint as to why some firms appear to resist organizational change, despite its high potential returns. For high initial levels of productivity, technical evolutions create opposite incentives for the decision-maker and for the upper layers of agents (managers).*

[11] Note that we are working in a partial equilibrium framework: w is kept exogenous in the model. However, our results should rather be reinforced if we introduced general equilibrium concerns: empirically, higher productivity increases wages. Since, from Proposition 1, higher wage costs reduce delegation and employment, employment should be further reduced in the organization.

To understand this, imagine a situation in which agents value positively the number of subordinates and layers beneath them, for a reason that is left out of the present analysis. In that case, ***for low initial levels of productivity****, both the decision-maker and the middle management (upper tiers in the hierarchy) will support the implementation of technical advances and their concomitant organizational changes: the decision-maker increases her profits, and higher tiers in the hierarchy get to control more subordinates.*

By contrast, ***for high initial levels of productivity****, technical progress implies that the number of layers and of agents should be reduced. In that case, upper tiers in the hierarchy will oppose the implementation of technical advances, since it would require them to reduce the number of their subordinates as well as the number of layers they control. In this case, if upper layers have some control on the shape of the organization, we should observe the organizational structure to adapt sluggishly when technical/organizational innovations become available, despite the existence of complementarities between higher productivity and organizational change.*

5 Conclusions

6 Appendix: Just-in-time Processing

According to the results of the previous section, enhancements of the agent's ability or capacity to make decisions can explain most of the observed evolutions in the organization's structure. In reality, more changes occurred. Different modes of organizing and coordinating work became possible. Lean production, just-in-time, instant communication, U-form organizations, have deeply modified the way in which organizations work. One may thus think that, in essence, it was not the agents' ability or capacity that lied at the heart of corporate evolutions, but instead the switch from some type of "sequential processing" to another coordination method.

In this section, we explore this possibility by considering a different way of coordinating the workers' tasks: just-in-time. The presumption is that the switch from one mode of organization to another simply happens or not. Instead, within each mode of organization, incremental improvements keep taking place. Therefore, if we reach the conclusion that such incremental improvements generate the same organizational dynamics as under sequential processing, it should not be the switch itself that generates the long-term pattern of organizational restructuring (hierarchization or flattening). Yet, the switch may have *other* effects, that can be highlighted by comparing the optimal organization under each mode of organization.

We choose to analyze just-in-time procedures for another reason as well. As stressed in Section 2, we assumed that agents in a layer l await the outcome of their subordinates' work to initiate their task. However, in contrast with several models of information processing such as Radner (1993) and Bolton and Dewatripont (1994), we assumed away "skip-level reporting". This allowed us to obtain tractable solutions for the optimal shape of the organization. A drawback is that all layers but one are kept idle when the latter layer is busy. Just-in-time procedures correct this organizational flaw, and will allow us to check whether our previous results are robust to other modes of organization.

Formally, we assume that, to reduce the amount of time a given layer is kept idle, each agent i processes his mass of information m_i in k steps. As soon as he completes the processing of a step (a fraction $1/k$ of his batch m_i), he transmits the output of his work to his superior, and starts working on the next step of his batch.[12]

[12] In our continuous-information setup, allowing explicitly for skip-level reporting makes the eventual math-

In this setup, *just-in-time* modes of organization can be represented as the limit $k \to \infty$, such that all layers start working at the same moment in time. Clearly, just-in-time is not always feasible in practice. There exist cases in which an agent must be aware of many (if not all) elements of information to make a decision (think for instance of information complementarities). When this is the case, the "sequential" organization makes perfect sense: all agents in a layer process their own elements of information, meet together, make their decisions and thereby discard a fraction $(1 - f)$ of the projects on the table, and then transmit the undiscarded items (unresolved tasks) to their superiors. Conversely, there also exist cases in which each decision can be made with very limited information regarding the other elements in the set M. These latter cases provide the best candidates to implement a just-in-time organization.

6.1 Just-in-Time Technology and Delay

Within a just-in-time organization, each worker transmits the output of his work to his superior on an item-per-item basis. When this way of organizing work is feasible, it generates shorter delays than the former "sequential" type of organization, since workers in a layer l do not have to wait that their subordinates have completed their task before initiating their own work. Instead, under just-in-time procedures, all layers initiate their work at approximately the same moment. Whenever feasible, the decision-maker will thus implement this mode of organization.

How does just-in-time affect our results? First, the definition of aggregate delay is modified. Under just-in-time, information flows in a continuous fashion from one layer to another. The flow at which the organization can process information is thus limited by the narrowest bottleneck in the organization; that is, by the pace at which the *slowest* layer processes information.

The delay needed to process a mass M of information items is thus *equal* to the time needed by the slowest layer in the organization, instead of the sum of each layer's delays:

$$D^{JiT}(L, \{n_l\}) = M \times \max_{l \geq 0} \left[\frac{f^{L-l}}{n_l} \right] + \phi \sum_{l=1}^{L} n_l, \text{ with } n_0 = 1. \tag{12}$$

To get a better grasp of (12) consider for a moment the case in which the organization is composed of exactly one layer ($L = 1$). If there is exactly 1 agent in that layer, he is alone

ematical expressions much more cumbersome. This is why we do not consider such a mode of organization in our model.

to process all M items. He needs M units of time to perform that task. In her stead, the decision-maker processes fM items, which requires fM units of time. Layer 1 is thus the bottleneck in this organization. It keeps the decision-maker idle every fraction $(1-f)$ of time, and total delay is only determined by that slowest layer: $D = M + \phi$.

To correct the situation, the decision-maker could hire more agents in layer 1. They would work in parallel and complete their task within a delay M/n_1. Increasing n_1 can thus speed up the delay needed by the organization, up to the point in which the bottleneck becomes the decision-maker herself:

$$D^{JiT}(1, n_1) = M \times \max\left[\frac{1}{n_1},\, f\right] + \phi n_1.$$

For that reason, increasing $n_1$above $1/f$ can only increase total delay: it increases coordination costs, and is thus counterproductive.

The optimization problem related to such bottlenecks remains similar when more than one layer are present in the organization: the decision-maker must make sure that the allocation of workers is optimal across these layers. Assume for a moment that the decision-maker has N agents in her organization. How should they be allocated across layers? Take two of them: layers l and j. Layer l is slower if f^{L-l}/n_l is larger than f^{L-j}/n_j. If this is the case, reallocating some workers from layer j to layer l can only reduce total delay.

Applying that reasoning to all pairs of layers, one obtains that the optimal number of agents in any layer $l \geq 1$ should be made proportional to the optimal number of agents in another layer j: $n_l^* = n_j^*\, f^{j-l}$. Consequently, in an organization composed of L layers and N^{JiT} agents, total delay is minimal when:

$$N^{JiT} = n_L^* \sum_{l=1}^{L} f^{L-l} = n_L^*\ \frac{1-f^L}{1-f}, \text{ that is if: } n_l = n_L\ f^{L-l},\ \forall l \in \{1, ..., L\}. \tag{13}$$

This implies that total delay becomes:

$$D^{JiT}(L, n_L) = M\ \max\left[f^L, \frac{1}{n_L}\right] + \phi\ n_L\ \frac{1-f^L}{1-f}, \tag{14}$$

where the first member in the maximum operator represents the delay needed by the decision-maker, and the second member the processing delay needed by any layer of agents.

We have thus identified how the decision-maker can optimally allocate agents in an organization that is composed of a given number of layers and agents. It now remains to determine the optimal number of agents in an organization composed of L layers, and then

how many layers is optimal. This will identify the optimal organization, given the value of the parameters. Interestingly, we also found that, since N^{JiT} is a sufficient statistic to identify how many agents should be present in each layer (given N^{JiT}), the optimization decision involves the same number of variables as if the number of layers was 1. Summing up:

Lemma 2 *Under Just-in-Time, each agent in layers $0 < l < L$ must have a span of control equal to $1/f$, and processing delays must be equal across all layers of agents $l = 1, .., L$.*

It now remains to identify how the decision-maker can optimize her organization over N^{JiT} and over L. To this end, we derive the objective function of the decision-maker, and the way in which these costs vary with each of the two variables.

6.2 Optimal Hierarchies under Just-in-Time

To derive the objective function under just-in-time, it is useful to note that just-in-time leaves wage costs unchanged: they are still defined by (3) . Therefore, when the decision-maker has access to just-in-time modes of organization, her objective function becomes:

$$\min_{L,n_L} \; TC^{JiT} = r\left(M \; \max\left[f^L, \frac{1}{n_L}\right] + \phi \; n_L \; \frac{1-f^L}{1-f}\right) + wM\,\frac{1-f^L}{1-f}. \tag{15}$$

A consequence of the maximum operator present in this objective function is that the latter is only piecewise differentiable with respect to n_L:

$$\begin{aligned} \frac{dTC^{JiT}}{dn_L} &= -M/n_L^2 + \phi\frac{1-f^L}{1-f}, \; \forall \; n_L < f^{-L} \\ &= \phi\frac{1-f^L}{1-f}, \; \forall \; n_L > f^{-L}. \end{aligned} \tag{16}$$

One observes that, $\forall \; n_L > f^{-L}$, this derivative is strictly positive. Put differently, it cannot be optimal to increase n_L beyond f^{-L}: the organization's bottleneck turns out to be the decision-maker. Therefore, only two cases have to be considered. Either total costs are monotonically decreasing in n_L up to $n_L = f^{-L}$ or they reach a minimum for some value $\bar{n}_L < f^{-L}$. In the former case, the optimum number of agents must be exactly $n_L^* = f^{-L}$: the number of agents is adjusted to exactly match the flow of the decision-maker. In the latter case, the first order condition:

$$\frac{d}{dn_L} r\left(\frac{M}{n_L} + \phi \; n_L \; \frac{1-f^L}{1-f}\right) = 0 \quad \Leftrightarrow \quad \bar{n}_L = \sqrt{\frac{M}{\phi}\frac{1-f}{1-f^L}}, \tag{17}$$

identifies the optimal number of agents. From (16), one finds that $\bar{n}_L < f^{-L}$ if and only if:

$$\begin{aligned} L > \bar{L} &\equiv \frac{\log\left[-1+\sqrt{1+4(1-f)M/\phi}\right] - \log\left[2(1-f)M/\phi\right]}{\log[f]} \\ &= \frac{\log 2 - \log\left[1+\sqrt{1+4(1-f)M/\phi}\right]}{\log f}, \end{aligned}$$

that is, if the number of layer is larger than this threshold $\bar{L}$.[13] Expressed differently, the optimal number of agents is found to exclusively depend on the number of layers in the organization. The intuition for this result is relatively straightforward. Since processing delays must be equalized across the layers $l = 1, .., L$, if one more agent is hired in a layer (and hence processing delay is reduced in that layer), he is unproductive unless additional agents are hired in all other layers. Hence, the larger is L, the "costlier" (in terms of coordination costs) it is to speed up processing. For $L < \bar{L}$, this aggregate coordination cost is "small", and in equilibrium the bottleneck is shown to be the decision-maker. This is why $n_L^* = f^{-L}$ in that case. Conversely, when the number of layers is larger than $\bar{L}$, coordination costs are "large" in the sense that they dominate the benefits of making the agents process information as fast as the decision-maker. Hence:

$$n_L^* = \begin{cases} f^{-L}, \ \forall L < \bar{L} \\ \\ \sqrt{\frac{M}{\phi}\ \frac{1-f}{1-f^L}}, \ \forall L \geq \bar{L}. \end{cases} \tag{18}$$

Using this result, we can now express total costs as a function of the number of layers only:

Lemma 3 *The lowest feasible cost associated with a just-in-time organization composed of L layers is given by:*

$$\begin{aligned} TC_1^{JiT} &= r\left(M\, f^L + \phi\ f^{-L}\ \tfrac{1-f^L}{1-f}\right) + wM\ \tfrac{1-f^L}{1-f}, \quad \forall L < \bar{L}, \ \textit{and} \\ TC_2^{JiT} &= 2r\sqrt{\phi M \tfrac{1-f^L}{1-f}} + wM\ \tfrac{1-f^L}{1-f}, \qquad\qquad \forall L \geq \bar{L}. \end{aligned}$$

Proof. *Immediate from (15) and* (18). ■

Using this lemma, we can evaluate how total costs vary with the number of layers. We find that TC_2^{JiT} is always increasing in L:

$$\frac{\partial TC_2^{JiT}}{\partial L} = -\ln(f)\,\frac{Mf^L}{1-f}\left(\sqrt{\phi/M\frac{1-f}{1-f^L}}\ r + w\right) > 0.$$

[13] One can easily check that this threshold is always positive, and monotonically increasing in f.

Therefore, it cannot be optimal to increase L beyond $\bar{L}$. The optimal number of layers is thus the value of L that minimizes TC_1^{JiT}, subject to the constraint that L should not be larger than $\bar{L}$:

$$L^* = \arg\min_L \left[TC_1^{JiT} + \lambda \ \left(\bar{L} - L\right)\right],$$

where λ is the Lagrangian multiplier associated with the constraint. Two trade-offs have emerged. On the one hand, the minimization of TC_1^{JiT} involves the same trade-off as under sequential processing: while additional layers help reduce processing delays, they increase coordination and wage costs. On the other hand, an additional constraint appears under just-in-time: the speed at which the agents process information cannot be higher than that of the decision-maker. The combination of these two trade-offs determines the optimal shape of the just-in-time organization:

Proposition 3 *Under just-in-time, the optimal organization is such that:*
a) all agents work the same amount of time as the decision-maker;
b) the optimal number of layers is given by:

$$L^* \in \{\lfloor \Omega \rfloor \,;\ \lceil \Omega \rceil\}, \text{ where } \Omega = \max\{0, \tfrac{\log(\phi/M) - \log[Z]}{2\log[f]}\},$$

where Ω is strictly decreasing in w/r and in ϕ/M, and is hump-shaped in $(1-f)$;
c) the span of control of both the decision-maker (n_1^{**}) and of the agents $\left(n_{l+1}^{**}/n_l^{**}\right)$ is $1/f$.

Proof. a) *Since, at the optimum, one must have $L^* \leq \bar{L}$, we find: $n_l^* = 1/f^{-l}$, $\forall l \in \{1, ..., L^*\}$. From (15), this implies that each layer of agents processes information during an amount of time equal to: Mf^{L^*}, and so does the decision-maker.*

b) *We proceed in two steps to prove part* b. *First, we look for the value L^{**} that minimizes TC_1^{JiT}, abstracting from the constraint $L \leq \bar{L}$. Second, we show that this constraint is never binding (as long as L is not constrained to be an integer number).*

Differentiating TC_1^{JiT} with respect to L obtains:

$$\begin{aligned} \frac{dTC_1^{JiT}}{dL} &= Mf^L \ \log[f] \ \left(r - \frac{w}{1-f}\right) - r\phi \ f^{-L}\frac{\log[f]}{1-f} = 0, \ \textit{or:} \\ f^{2L} &= \frac{\phi}{M}\left(1 - f - \frac{w}{r}\right)^{-1} = \frac{\phi}{M} \ Z^{-1}, \ \textit{with } Z = (1-f) - w/r. \end{aligned}$$

Taking logarithms yields:

$$\Omega = \max[0, \frac{\log(\phi/M) - \log[Z]}{2\log[f]}].$$

Now, we show that $\Omega < \bar{L}$ *for* $w/r = 0$:

$$\Omega = \frac{\log(\phi/M) - \log[1-f]}{2\log[f]} < \frac{\log 2 - \log\left[1+\sqrt{1+4(1-f)M/\phi}\right]}{\log f} = \bar{L},$$

which is true if and only if:

$$\log(\phi/M) - \log[1-f] > \log 4 - 2\log\left[1+\sqrt{1+4(1-f)M/\phi}\right]$$
$$\log 4 < \log\left[\phi^{\frac{2+4(1-f)M/\phi+2\sqrt{1+4(1-f)M/\phi}}{M(1-f)}}\right] = \log\left[4+2\phi^{\frac{1+\sqrt{1+4(1-f)M/\phi}}{M(1-f)}}\right],$$

which always holds. Therefore, we proved that Ω *is always strictly smaller than* $\bar{L}$ *in* $w/r = 0$, *. Next, it is easy to check that* $dL^{**}/d(w/r) < 0$, *while* $d\bar{L}/d(w/r) = 0$. *Hence:* $\Omega < \bar{L}$ *for any admissible set of parameters, which proves that the constraint* $L < \bar{L}$ *cannot be binding in equilibrium (this result does however not take integer constraints into account).*

Differentiating Ω *with respect to* f, *one finds that:*

$$\frac{d\Omega}{df} > 0 \Leftrightarrow 2\Omega\left(1-f-\frac{w}{r}\right) > f,$$

and hence that $d\Omega/df$ *is positive for* $f \to 0$ *but negative for* $f \to 1 - w/r$. ■

The comparison between the sequential and just-in-time modes of organization reveals that most of our previous results remain valid. In particular, technical progress (an increase in $1-f$) induces the decision-maker to increase the number of layers in her organization if f is initially close to 1, and to implement "delayering" if f is already low. The main differences between the two types of organization are that 1) total delay is shorter under just-in-time (wage costs remain unchanged). Therefore, there exist sets of parameters for which the decision-maker would rather work alone than set up a *sequential* organization, and rather set up a *just-in-time* organization rather than work alone. 2) Just-in-time introduces a new "bottleneck" constraint on the delay needed by the agents to process information: this delay cannot be shorter than that of the decision-maker. As a result, the span of control of the decision-maker is also bound upwards. For f sufficiently large (respectively: f sufficiently small), introducing just-in-time in an initially sequential organization thus tends to reduce (resp: increase) the optimal span of control of the decision-maker. 3) The same bottleneck constraint generates an upper limit on the number of layers in the organization. While this limit is never binding at the optimum, it may become binding when the number of layers is forced to be an integer number. By means of numerical simulations, we indeed found that this constraint is actually binding for some combinations of parameter values, and this is why

the optimal (integer) number of layers can either be the upper or the lower integer number closest to Ω, depending on parameter values.

Another noteworthy difference between the two modes of organization stems from the constraints that just-in-time imposes on the delay needed by each layer to process information:

Proposition 4 *Under just-in-time, there is thus full complementarity between technical and organizational change. That is, in the absence of organizational restructuring, technical progress does not generate* ***any*** *reduction in delay.*

Proof. *For a given L and n_L, delay is given by: M/n_L. Thus, unless n_L and/or L are re-optimized, total delay cannot change.* ■

The complementarity between the level of the technology and the shape of the organization was already apparent under sequential processing, since the optimal organization was a function of the agents' productivity. Yet, just-in-time imposes another constraint on processing delay in each layer, which makes this complementarity even more salient. When technical progress increases, if the number of agents remains unchanged in the lowest layers of the organization, there cannot be *any* reduction in delay. Taking advantage of technical progress requires that the shape of the organization be adapted. This is exactly the puzzle highlighted by the productivity paradox: IT services were spreading fast across industries. Yet, measured productivity has not been increasing at all for many years. BBH, Brynjolfsson and Hitt (2000), or Rajan and Wulf (2006) instead suggest that such productivity gains only materialized when the structure of the organization could be adapted meaningfully.

Depending on the cases, adapting the shape of the organization may either call for an expansion or for a reduction in the number of agents and/or number of layers. If, for some reason, the decision-maker is constrained to hold fixed the total number of agents in the organization, she would have to demote some agents; reallocate them from a position high up in the hierarchy to the lowest layers. If there is resistance to such a change (See Remark 1), she would have to hire more agents, essentially for these lowest layers, even though some agents higher in the hierarchy may remain idle part of the time. Clearly, this constrained reoptimization may go in the opposite direction to what is optimal, *i.e.* to reduce the overall number of agents and of layers when f is not too large. Resistance to organizational restructuring may thus prove costlier under just-in-time than under a sequential mode of organization.

References

[1] Acemoglu, D. 1998. Why Do New Technologies Complement Skills? Directed Technical Change and Wage Inequality. *Quarterly Journal of Economics*, 113(4): 1055-89

[2] Aghion, P. and P. Howitt 1997. *Endogenous Growth Theory.* Cambridge: MIT Press

[3] Autor, David, Lawrence Katz, and Alan Krueger (1998) "Computing inequality: Have Computers Changed the Labor Market?", *Quarterly Journal of Economics* 113(**XXX**): 1169-1213.

[4] Bolton, P. and M. Dewatripont 1994: The Firm as a Communication Network. *Quarterly Journal of Economics* 109(4): 809-839.

[5] Bresnahan, T., E. Brynjolfsson and L. Hitt 2002: Information technology, workplace organization, and the demand for skilled labor: Firms Level evidence. *Quarterly Journal of Economics* : 339-376.

[6] Brynjolfsson, Erik and Lorin M. Hitt (2000) "Beyond Computation: Information Technology, Organizational Transformation and Business Performance." *Journal of Economic Perspectives*, 14(4): 23-48.

[7] Brynjolfsson, Erik and Lorin M. Hitt (2003) "Computing Productivity: Firm-Level Evidence," *Review of Economics and Statistics* 85(4): 793-808.

[8] Castanheira, M. and M. Leppämäki 2003: Communication and Information Management: Organizations and Markets. *CEPR Discussion Paper* No. 4072.

[9] Caroli, Eve and John Van Reenen (2001) "Skill-biased organizational change? Evidence from a panel of British and French Establishments", *Quarterly Journal of Economics* **XXX**: 1449-1492.

[10] Crémer, Garicano and Prat 2005: Codes in Organizations. Toulouse mimeo.

[11] Domberger, S. 1998: The Contracting Organization. A Stregic Guide to Outsourcing. Oxfrod: Oxford University Press.

[12] Garicano, L. 2000: Hierarchies and The Organization of Knowledge in Production. *Journal of Political Economy* 108: 874-904.

[13] Garicano, L. and E. Rossi-Hansberg 2003: Organization and Inequality in a Knowledge Economy. mimeo. University of Chicago

[14] Grossman, G. and E. Helpman 1991: *Innovation and Growth in the Global Economy.* Cambridge: MIT Press

[15] Grossman, G. and E. Helpman 2002, "Integration vs. outsourcing in industry equilibrium", *Quarterly Journal of Economics* 117(1): 85-120

[16] Meagher, K., H. Orbay and T. Van Zandt 2001: *Hierarchy Size and Environmental Uncertainty.* INSEAD mimeo

[17] Meagher, K. 2003: Generalizing incentives and loss of control in an optimal hiearchy: the role of information technology. *Economics Letters* 78. 273-280.

[18] Prat, A. 1997: Hierarchies of Processors with Endogenous Capacity. *Journal of Economic Theory*, 77: 214-222.

[19] Qian, Y. 1994: Incentives and Loss of Control on an Optimal Hiearchy. *Review of Economic Studies* 61. 527-544

[20] Radner, R. 1993: The Organization of Decentralized Information Processing. *Econometrica*, 62: 1109-1146.

[21] Radner, R. and T. Van Zandt 1992: Information Processing in Firms and Returns to Scale. *Annales d'Economie et de Statistique* 25/26: 265-298

[22] Rajan, R. and J. Wulff 2006: The Flattening Firm: Evidence from panel Data on the Changing Nature of Corporate Hierarchies. *The review of Economics and Statistics* 88: 759-773

[23] Sah, R.K. and J. Stiglitz (1986). "The Architecture of Economic Systems: Hierarchies and Polyarchies." *American Economic Review* 76(4). 716-727

[24] Wulf, J. 2012: The Flattened Firm: Not As Advertised. *California Management Review* 55. 5-23.