Organizational Economics with Cognitive Costs*

Luis Garicano and Andrea Prat
London School of Economics

March 2011

Abstract

Organizational economics has advanced along two parallel tracks, one concerned with motivating agents with diverging objectives, the other – less developed – with coordinating agents under cognitive limits. This survey focuses on the second strand and attempts to bring the two strands together. Organizations are viewed as responses to the cognitive costs faced by their (potential) members. We review existing approaches such as team theory, hierarchies of processors, organizational languages and knowledge hierarchies and we argue that they can help us address an array of important organizational issues. We also review recent developments in the application of these ideas: exploiting complexity measures, combining team theory and contract theory, applying organization theories in labor economics, and using these theories to interpret the wealth of activity data that is becoming available.

Keywords: bounded rationality, tacit knowledge, codifiable knowledge, code, vertical communication, horizontal communication, organizational architecture, decision bias

JEL Classifications: D2, D8, L2, M5

*We thank Daron Acemoglu, Michael Best, Jacques Crémer, Glenn Ellison, Bob Gibbons, Navin Kartik, and audience participants at the Econometric Society World Congress in Shanghai and MIT for useful comments.

1 Introduction

Organizations are formed by agents working together for a common purpose. Two sets of obstacles may prevent organizations from reaching their goals. First, incentive problems: achieving a common goal may be difficult because individual pursue different objectives. Second, even when all agents share a common purpose, they face coordination problems due to cognitive costs. They must exchange information, coordinate their actions, and make joint decisions. Such activities require talking, writing, reading, and thinking – all tasks that require time and energy.

On the first set of problems, starting with the work of Leo Hurwicz (1973) and Theodore Groves (1973), mechanism design and contract theory have made huge progress. Contracts are designed, under asymmetric information, so that principals can motivate their agents in order to align their individual goals with those of the organization. The typical contract-theoretic contribution strives to find the optimal contract, holding constant the role of the agent in the organization and the informational environment. For instance, in the classical moral hazard problem, the set of possible actions that the agent can take is given and so is the information that the agent and the principal observe; moreover, the relation between agents is also exogenously given. This corresponds to assuming an exogenous solution to the second problem discussed above.

However, in practice organizations can choose, at least partly, how to allocate tasks to agents, what monitoring systems to put in place and what channels of communication to establish. In a world with cognitive limits, these organizational choices come at a cost. Designing the organization requires choosing between alternative structures given the cognitive limits of its agents (or the agents that could be hired).[1]

Economists have been aware of the importance of bounded rationality for the study of organizations for a long time. Simon (1947) and March and Simon (1958) advocated studying "the motivational, conflict of interest, cognitive and computational constraints that human beings place on organizations." In particular they emphasized the importance of the role of individual cognition in understanding organizations and "the constraints placed on the human being by his limitations as a complex information-processing system." In Simon's view, "organization members are decision makers and problem solvers and [...] perception and thought processes are central to the explanation of behavior in organizations" (March and Simon, 1958, p 25).

Arrow (1974) is a manifesto for a theory of organizations built on bounded rationality. He argues that since individuals are boundedly rational, an organization can acquire more information than any individual, and thus an organization is useful for information handling.

[1]The two main elements of this problem – the presence of important cognitive costs and the endogeneity of the organizational solution – were identified by Hayek (1945). In his view, the key problem that a society faces is that of aggregating dispersed information: "Knowledge of the circumstances of which we must make use never exists in a concentrated or integrated form, but solely as the dispersed bits of incomplete and often contradictory knowledge which all separate individuals possess."

In particular, he highlights two advantages of organizations: first, since communication channels must be designed, organizations can choose a specialized code for efficiency; second, since joint production is preferred, but exchanging information is costly, authority is useful because it economizes on information costs.

The goal of this survey is to review the economic literature on organizational responses to bounded rationality. We include contributions that satisfy three criteria: (i) Cognitive limits are part of the model (e.g. agents face a cost of communication); (ii) They deal with issues of interest to organizational economists, such as task allocation or the creation of organizational languages; (iii) The structure of the organization is, at least to a certain extent, chosen in response to the cognitive limits of its members (e.g. tasks are allocated to reduce the need for communication).

On the methodological side, we are inspired by Herbert Simon's emphasis on being explicit about cognitive assumptions. We categorize the existing literature on the basis of the constraints that makes coordination costly or impossible.

In section 2 we begin with the model that has guided scholars in this area since the 50s: team theory. Team theory studies multi-agent problems where agents share the same goal but may have asymmetric information. After a short presentation of the theory, we review a number of applications, which shed light on important organizational phenomena. The material is structured around substantive questions such as "How should functions be grouped into divisions?" or "Should teams be homogenous?"

Team theory supplies a set of powerful tools to characterize the benefits of different information structures. We can, for instance, ask whether the organization is better off if a particular agent receives information from a certain source or from a different source. However, team theory is silent on the cost of different information structures. Hence, it is usually complemented by ad hoc assumptions on the cost and feasibility of different information structures.

Since the 90s, a small but growing number of scholars have been exploring the cost side as well. The approach, which we review in section 3, consists in modeling a particular form of cognitive cost in detail and studying how organizations would structure themselves in response to this cost factor. We identify three sets of models, according to whether the cognitive limit concerns the variable cost of information processing and communication, the fixed cost of communication, and the cost of knowledge. The three cognitive costs give rise respectively to three sets of models: hierarchies of processors (Radner 1993, Bolton and Dewatripont 1994), organizational languages (Crémer, Garicano and Prat 2007), and knowledge hierarchies (Garicano 2000).

Section 4 discusses recent efforts by scholars to incorporate both incentive and cognitive considerations in the design of firm's organizations. It reviews a number of papers that combine team theory and contract theory to draw important lessons on the interaction between incentive structures and coordination mechanisms.

Section 5 reviews some current developments of the literature. First, at a foundational level, there are attempts to use complexity notions to microfound cognitive costs. As computer scientists have developed a sophisticated body of knowledge on complexity, such a linkage would provide powerful foundations to organization theory. Second, a stream of research looks at the interaction between the labor market and the internal organization of firms that have resulted from advances in information and communication technology and in the international division of labor. Third, organization theories based on cognitive costs yield testable implications in terms of how agents allocate their time to different tasks. Thanks to electronic records, activity data is becoming more readily available and provides organizational economists with a chance to use the intellectual framework they have been developing to understand how organizations work in practice.

2 Team Theory

Team theory is a powerful and general tool for economists working on organizations. It represents the natural extension of single-agent decision making under uncertainty to multiple agents. Agents share the same objective but they observe different signals. When they make decisions, they have only partial information on what their team members have observed. In our categorization, team theory relates to limits to communication. Implicitly, agents have different information sets because they cannot fully communicate with each other.

Team theory was developed starting in the 50s by Jacob Marschak and Roy Radner (see Marschak and Radner 1972 for the most complete treatment). Later, it was somewhat neglected as the interest of the profession shifted to noncooperative games. However, in recent years economists have used team theory to derive a number of insights in organization economics. Moreover, some team-theoretic techniques have been increasingly applied to game-theoretic problems.

The general lesson that emerges from the team theory literature is that there are important strategic complementarities between the information structure of a firm and other organizational variables such as decision rules, workforce characteristics, delegation, and task specialization. Team theory allows us to make a rich set of predictions about these complementarities.

2.1 Decision-making in teams: Basic framework

Let X be the set of possible states of the world with an associated probability distribution π. Let A_i be the set of actions available to team member i. The team payoff is

$$\omega\left(x, a_1, \ldots, a_n\right)$$

where x is the realized state and a_i is the action chosen by member i. Implicit in this set-up is the assumption that team members pursue the same objective, namely the maximization of the expected value of the payoff ω.

Typically, agents do not observe the state of the world directly. Let

$$y_i = \eta_i(x)$$

be the signal that i receives when the state of the world is x. The function η_i is i's *information structure*. Member i makes a decision based on his own signal realization

$$a_i = \alpha_i(y_i).$$

The function α_i is i's *decision rule.* Given an information function and a decision rule for every agent, the team's expected payoff is

$$\Omega(\alpha, \eta) = E[\omega(x, \alpha_1(\eta_1(x)), \ldots, \alpha_n(\eta_n(x)))].$$

With these concepts in mind, it is easy to see the role of team theory within organization economics. The typical way of proceeding is to fix the information structure of agents and to determine the optimal decision rule together with the maximum payoff

$$\hat{\Omega}(\eta) = \arg\max_{\alpha} \Omega(\alpha, \eta).$$

The optimized payoffs can then be used to compare different information structures. If we wish to determine whether information structure η'' is better than information structure η' we just need to compare $\hat{\Omega}(\eta'')$ to $\hat{\Omega}(\eta')$. One can also imagine that the two information structures have different costs $c(\eta')$ and $c(\eta'')$, in which case the correct comparison will be $\hat{\Omega}(\eta'') - c(\eta'')$ to $\hat{\Omega}(\eta') - c(\eta')$.

Team theory was developed before Harsanyi's analysis of games of incomplete information. We can now view the set-up just described as a special case of a static game of incomplete – one where all agents have the same payoff function ω. The information structures are common knowledge and y_i is the type of player i. Decision rules are interpreted as strategies.

The commonality of interest among players implies two results that are not typically true in game-theoretic settings. First, an increase in the accuracy of the information (in a Blackwell sense) that a certain agent i receives cannot decrease the expected payoff. This is immediate because the agent could still use the same decision function he used before, so he would not move to a new decision function if it yields a lower expected payoff for the team.

Second, a necessary condition for the payoff to be maximized for all members is that it is maximized for each member individually. For every i, holding fixed the α's of the other

members at the optimal level α^*, it must be true that

$$\hat{\Omega}(\eta) = \arg\max_{\alpha_i} \Omega(\alpha_1^*, \ldots, \alpha_i, \ldots, \alpha_n^*, \eta).$$

This fact, labelled by Marschak and Radner as *person-by-person optimality*, is not necessarily true in general games, where Pareto-efficient outcomes need not be supported by Nash equilibria. If the payoff function is concave and differentiable in actions, then person-by-person optimality is both necessary and sufficient for optimality – a fact that greatly simplifies the search for the optimal decision rule.

For economists interested in organizations, team theory has at the same time a great strength and a great weakness.

The strength of team theory has to do with the generality of the approach, which can accommodate an enormous variety of information structures. To illustrate this point, let us consider *management by exception*, namely the idea that workers act independently in normal situations while they call an emergency meeting whenever one of them observes something abnormal. Management by exception can be modelled in a simple way within team theory. Let μ_i denote the information structure of member i if he acts independently and let R_i be the set of emergency signals. Then, we can write the final information structure of member i as follows

$$\eta_i(x) = \begin{cases} \mu_i(x) & \text{if } \mu_j(x) \notin R_j \text{ for all } j = 1, \ldots, n \\ \{\mu_j(x)\}_{j=1,\ldots,n} & \text{if } \mu_j(x) \in R_j \text{ for at least one } j = 1, \ldots, n \end{cases}$$

One can then assume that holding an emergency conference has a fixed cost F and can determine the optimal set of 'exceptional' signals (Marschak and Radner 1972, Chapter 6).

In fact, team theory can even be extended beyond the one-shot environment described above. Marschak and Radner (1973, Chapters 7 and 8). For instance, we can examine communication protocols whereby Agent 1 passes a possibly imperfect signal to Agent 2, who combines it with his own signal and passes it to 3, and so on.

The main weakness of team theory is related to its strength. It allows comparison of the benefits of a wide range of information structures, but it says nothing about their cost or feasibility, which are determined outside the model. Hence, we are left with a large number of potential solutions. Team theory must be complemented by assumptions based on the observation of real organizations if we want to derive precise predictions.

2.2 Applications of Team Theory

There are a number of team-theoretic contributions that yield important insights for organizational economists. Here we review some of these results with the objective of illustrating the empirical implications that the analysis yields, and which, in our view, serve to illuminate questions on which more motivation/incentive based models have been silent.

Question 1. How should functions be grouped? Crémer (1980) asks how functions within a firm should be optimally grouped into divisions. He considers a firm composed of many basic units, each of which is subject to some source of uncertainty which it observes. Specifically, suppose that x_{ik} is the production by unit i of good or service k, $k = 1, 2, ..., m$, and $C_i(x_i)$ the cost of producing the vector $x_i = (x_{i1}, ...x_{im})$. for each good k, total production must equal demand x_k^d. Letting x be the production matrix of all units, the full-information problem is simply:

$$\min C(x) \quad \text{s.t. for all } k, \sum_{i=1}^{n} x_{ik} = x_k^d.$$

Suppose, however, that each unit is subject to cost uncertainty. Units must be organized in division. A division observes the shocks that affect its units but not the units belonging to other divisions. Equivalently, transfers between divisions are set before uncertainty is realized, while intra-divisions transfers take place after the resolution of uncertainty.

Crémer shows that the solution is a partition of the set of units that *minimizes the variance of transfers* between divisions. In other words, a good organizational structures groups together functions in a way that makes the interaction between divisions as predictable as possible. Crémer's result helps us identify organizational dilemmas. These are situations where there is no simple way of eliminating inter-division variance.

Question 2. How should managerial time and talent be allocated? Geanakoplos and Milgrom (1991) expand the analysis in Crémer (1980) by introducing managers. They study a hierarchy in which managers' role is to increase the amount of information brought to bear on each particular decision, but managers are limited in the amount of information they can collect. The questions they pose include: What are the key trade-offs determining the hierarchy structure? How is the organization affected by uncertainty in the environment?

The objective of the organization is to allocate resources among units. The inputs and outputs are given by a multidimensional vector (we will do it in one dimension); the cost function is simply a quadratic form of those vectors. The problem is that the intercept of the cost function is unknown, only the slope is known. The managerial role is to decide how much to allocate to each unit, but only the allocation can be communicated. Each manager is told by her superior how many resources she gets, and tells that to subordinates. Managers aim to minimize expected total costs of their units.

Let γ_i be the cost intercept. Managers have limited time: if a manager of ability α_i spends time τ_i studying unit i, then observes: $\gamma_i + \frac{\varepsilon_i}{\alpha_i \tau_i}$, with $\varepsilon_i \rightsquigarrow N(0, s)$. The role of multiple managers is to bring more time to bear on decisions. The solution is obtained exactly in a team theoretical form: given information structure and organization, solve for the resource allocation; and then optimize over information structure given organization.

Among the implications of the analysis, they find that the number of managers is de-

creasing in managerial wage, is increasing in the value of information, and follows an inverted U-shaped with respect to managerial ability. Intuitively, when managers are unproductive, firms reduce number of managers to avoid wage costs; when they are productive, all possible savings can be reached by few managers.

Question 3. Should teams be homogeneous? Crémer (1993) studies whether firms should have shared specific knowledge (e.g. of facts, of rules of language). His answer is that this should be the case if actions are complementary. Essentially, agents see the state of the world plus a disturbance. If coordination is important, diversified information will dampen the reaction to information about state; while, if coordination is of little importance, diversified information makes it feasible for the agents to react strongly to the information that they receive.

Prat (2002) expands on the analysis by asking whether team members should be homogeneous or heterogeneous. Suppose a firm can potentially hire two types of workers – at the same wage. The workers are equally able and motivated but they 'see' the world differently. Under what conditions would the firm prefer a homogenous workforce? In practice, we observe very different answers to this question, ranging from military organizations, which typically require all their members to attend the same schools, to innovative businesses, which emphasize the value of workforce diversity.

In team-theoretic terms, suppose the two types of workers are associated with information structure η' and η''. A *homogeneous configuration* is either one where all agents have η' or one where they all have η''. A *heterogenous configuration* is any situation where some agents have η' and others have η''.

Finally, the team payoff function – assumed to be symmetric – is supermodular (submodular) if, for any two vectors $(a'_1, \ldots, a'_n)$ and $(a''_1, \ldots, a''_n)$,

$$\begin{aligned} & \omega\left(x, \min\left(a'_1, a''_1\right), \ldots, \min\left(a'_n, a''_n\right)\right) + \omega\left(x, \min\left(a'_1, a''_1\right), \ldots, \min\left(a'_n, a''_n\right)\right) \\ \geq & \ (\leq)\, \omega\left(x, a'_1, \ldots, a'_n\right) + \omega\left(x, a''_1, \ldots, a''_n\right). \end{aligned}$$

If f is twice differentiable, f is supermodular (submodular) if the second-order cross partial derivative between any two actions is everywhere positive (negative). We can then prove that if the payoff function ω is supermodular in the actions of team members, then the optimal configuration is homogenous. If it is submodular, then the optimal configuration is heterogenous.[2]

A supermodular payoff corresponds to a situation where agents' actions are complements. The payoff is greater if the actions that agents choose are positively correlated rather than uncorrelated or negatively correlated. Errors are less costly if they are correlated across

[2]For the submodular part of the result, it is necessary to assume that the two homogenous configurations yield the same expected payoff, otherwise it may be the case that having heterogeneous workers is not optimal simply because one type is much better than the other.

agents. Having a homogeneous configuration guarantees that agents receive the same signal and choose the same actions. Errors are perfectly correlated. The classical example of an organization that values coordination is the military. This proposition helps explain why officers tend to receive long, common training, which leads them to react to similar situations in similar ways.

On the contrary, with a submodular payoff, agents' actions are substitutes. The team prefers its members to make uncorrelated mistakes and that is why it opts for a heterogeneous configuration. Search problems tend to give rise to submodular payoff functions. In situations where experimentation is more important than coordination, such as innovation-based industries, we should expect organizations to value a diversity of backgrounds in the workforce they employ.[3]

Question 4. When should functional or divisional firms be preferred? Qian, Roland, and Xu (2000) consider a multi-unit firm, where some pairs of units face an attribute matching problem (e.g. a car engine must fit the car body) and other pairs face an attribute compatibility problem (e.g. using the same transmission system on all cars yields a cost saving). Two possible structures are considered: in the M-form, units are grouped according to product lines; in the U-form they are grouped by functions. As in Crémer (1980), units that are grouped together can share information more easily. The M-form facilitates attribute matching, while the U-form enhances attribute compatibility. Qian, Roland, and Xu characterize the terms of this trade-off. In particular, they show that the M-form is particularly useful when innovation is important as it allows for long-term small-scale experimentation. The reason is that in the M-form, local managers can solve the attribute matching problems among complementary tasks but are less capable of achieving attribute compatibility across the organization, while in U-form organization, local managers can solve attribute compatibility more easily.

Question 5. Centralization or decentralization? Error correction and robustness: In approving projects, allocating resources or launching new products, individuals make mistakes: they sometimes approve projects they should reject and they sometimes reject some they should accept. Organizations, by imposing extra checks on those decisions, may limit these errors, but they may introduce others. Taking that into account, organizations can be deliberately designed ex ante to complement limitations to human knowledge and judgement (Sah and Stiglitz 1986).

[3] Athey, Avery and Zemsky (2000) give a different answer to the question of how much homogeneity is desirable by introducing type specific mentoring: higher level workers increase the productivity of those of their own type. On the other hand, a more homogeneous firm loses out by passing on high quality people. Thus the trade-off is the quality of workers versus amount of mentoring. As the firm grows, can trace out different levels of homogeneity: if skill is scarce, firms care less about homogeneity; on the other hand, if skill is plentiful but mentoring important – then firms will be more homogeneous.

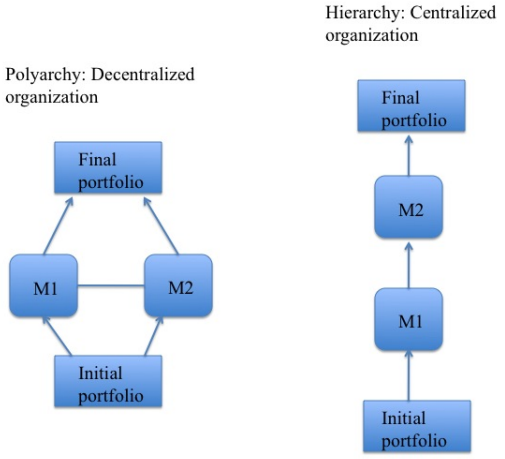

Figure 1: A poliarchy/market/decentralized architecture versus a hierarchy/centralized one

Agents in an organization need to evaluate and decide whether to accept a project. Limited rationality of the agents means that they make two types of mistakes when evaluating a project: reject a good project (Type-I error) or accept a bad project (Type-II error). With perfect screening, all good projects are accepted while the bad ones are rejected. Sah and Stiglitz (1986) consider two alternative screening systems: a polyarchy and a hierarchy (See Figure 1). In a polyarchy, decision making is decentralized to multiple independent screens. So a project is accepted if it is approved by any evaluator. In a hierarchy, decision making is centralized to the top through successive screens. A project is accepted only if it passes all evaluators' screening. Note that there is no communication in a polyarchy as all evaluators work independently and make their decisions simultaneously. In a hierarchy, the communication between the evaluators is limited to a binary signal "Accept", "Reject". Implicitly the evaluators' decisions are substitutes under polyarchy and complements under hierarchy. With two evaluators, the probability that a project is approved is $p^P(z) = s(z) + s(z)[1 - (s(z))] = 2s(z) - s^2(z)$ under polyarchy and $p^H(z) = s^2(z)$, where s is the screening function that determines the probability that a project with underlying value z 'looks' good and is accepted. Immediately we can see that with the same screening criterion a polyarchy is more likely to accept the project than a hierarchy. In other words, the incidence of making type-II error is relatively high under polyarchy while the incidence of type-I error is relatively high under hierarchy.

These ideas illuminate the stylized fact that well-established firms are not prolific at innovating. Well established firms want to protect their reputation and will be careful about the type of innovations they implement. The theory helps us illuminate how they do this: by establishing hierarchies, so that approving new products in a mature firm with a strong reputation will involve a highly bureaucratic process with numerous steps and procedures.

As a result, there is a high probability that good projects are rejected, and this is the price paid to avoid bad projects being accepted. Other examples may be an industry subject to a lot of public scrutiny or activities such as risk management where the loss is potentially large but gain is little.

Question 6. Adaptation or Coordination? Dessein and Santos (2006) use team theory to analyze the organizational trade-off between adapting to change and coordinating complementary activities. Task i involves choosing a vector of actions $a_i = (a_{i1}, \ldots a_{in})$. The first feature of the model is that a worker can handle more than one task. Tasks are partitioned across workers: assume for simplicity that all workers get the same number of tasks t. Let $T(i)$ denote the set of tasks assigned to the worker in charge of task i. Every task is subject to an independent local source of uncertainty, x_i, normally distributed with mean $\bar{x}_i$ and variance σ^2.

The team payoff is quadratic and takes the following form:

$$\omega(x, a_1, \ldots, a_n) = -\phi \sum_{i=1}^{n} (a_{ii} - x_i)^2 - \beta \sum_{i=1}^{n} \sum_{j \notin T(i)} (a_{ji} - a_{ii})^2 - h(t, \alpha).$$

The first element of the payoff function represents the benefits of adaptation. The primary action related to task i should be as close as possible to the local state x_i. The second element captures the benefits of coordination across tasks. The outcome of task i depends on ancillary actions relating to other tasks, which should fit as closely as possible with the primary action. The third element represents returns from specialization: the cost h is increasing in the number of tasks that a worker has to handle, t. The three parameters, ϕ, β, and α, measure respectively the importance of adaptation, coordination, and specialization.

If agents had full information, the problem would have a simple solution: assign each task to one agent and set $a_{ii} = a_{ji} = x_i$ for all i and all j. However, it is more reasonable to assume that agents can only observe directly the local uncertainty that relates to the tasks they are assigned to. The second key feature of Dessein and Santos is that it introduces an explicit communication stage. The agent in charge of task i tries to convey x_i to other agents. Transmission is successful with probability p. The sender does not know whether the message has gone through.

Given an organizational structure – defined by t – the optimal actions can be derived. The primary action for task i is

$$a_{ii}^* = \bar{x}_i + \frac{\phi}{\phi + \beta(n-t)(1-p)} (x_i - \bar{x}_i)$$

while the ancillary action linked to task j is

$$a^*_{ji} = \begin{cases} a^*_{ii} & \text{if communication is successful} \\ \bar{x}_i & \text{otherwise} \end{cases}$$

These expressions yield a clean characterization of the terms of the trade-off between adaptation and coordination. An increase in the importance of adaptation – higher ϕ – strengthens the link between the primary action a^*_{ii} and the local state. In contrast, an increase in the need for coordination – higher β – makes the primary action less variable. Better communication – a higher p – dampens the trade-off between adaptation and communication, allowing the primary action to track the local state and the ancillary action to track the primary action.

Task specialization, $\frac{1}{t}$ can now be endogenized, to maximize the expected organizational payoff. Dessein and Santos identify a strategic complementarity between the need for adaptation, job breadth, and local uncertainty. It is shown that task specialization is decreasing in the importance of adaptation ϕ and the variance of local circumstances σ^2. Having broad jobs facilitates coordination between primary and ancillary actions. Such coordination is most useful when primary actions are unpredictable, which in turn is due to the unpredictability of the local environment. Firms that face a lot of local uncertainty must forgo gains from specialization in order to achieve a better organizational response to the dual challenge of coordination and adaptation.

Dessein and Santos also endogenize the quality of communication p and they show that it is complementary to the need for adaptation, job breadth, and local uncertainty. The firm wants to invest more in communication precisely when it would like to reduce labor division. To sum up, one can identify two distinct organizational solutions: one based on labor division, limited communication, and inflexible actions; the other based on broadly defined jobs, communication investments, and high adaptation.

3 Microfounding Cognitive Costs

As we have seen, team theory can be used to analyze the benefit of different information structures, but it is silent about the underlying cognitive costs that make one information structure more or less expensive than another one. This methodology can still be used successfully to compare the benefits of different information structures, as we shall see shortly. However, as the information gathering process is not microfounded, the cost side of communication always has an ad hoc element.

In more recent years, there have been attempts to study organizations as optimal responses to explicit cognitive limits of their participants. Since the structure of the organization is, at least partly, an optimal response to the bounded rationality of the agents who work (or could work) for it, it is useful to categorize models in this area on the basis of the cognitive limits

they assume. The first approach – hierarchies of processors – offers a microfounded model of variable costs of communication and processing. Agents incur a cost every time they read or communicate a bit of information. The second approach – coding – instead models the fixed cost of communication. To exchange information more effectively, agents can create a shared organizational language. The cognitive cost is then measured by the complexity of the language that an agent has to learn. Finally, one can microfound problem-solving ability – hierarchies of knowledge – whereby agents must decide which pieces of knowledge to acquire at a cost that is increasing in the range of problems they can successfully tackle.

3.1 Endogenous communication: Hierarchies of processors

In our analysis of Geanakoplos and Milgrom's (1991) in the previous section we encountered an analysis of a decentralized hierarchy dealing with a resource allocation problem. There was no explicit information transfer beyond the allocation of resources by the superiors in the hierarchy to those below them. The role of decentralization was to bring more information to bear in decision making: since managerial time is limited, by creating a hierarchy we can have a manager who only allocates his attention to a subset of plants, and thus we can study those costs more carefully. The limitations on the information of each manager are, however, exogenous. An extensive literature, associated with Timothy van Zandt and Roy Radner among others, has endogenized communication and information aggregation in organizations from first principles, starting from the raw data up.

All papers in this strand of the literature start with a decomposable associative operation (such as adding up, or finding a maximum, among a set of numbers) and then allocates managers to process this problem in parallel. Agents in these networks read all of the primary data, or a report sent by others. In designing the network, the objectives may be to reduce delay, to minimize work load (the total time agents are busy), to minimize organizational size (number of agents) or to maximize throughput. The models generally rely on three parameters: the time it takes to perform associative operations, the time it takes to send data, and the time it takes to read data.

What are the costs and benefits of decentralization in terms of information processing? By processing data in a decentralized way, organizations reduce delay, increase the rate of processing output, and allow specialization in different types of tasks. The costs of such networks are that they increase communication costs, and administrators' wasted time.

To understand this approach consider Radner (1993), which sets up organizations to minimize delay, and assumes agents incur a fixed cost per message: each message can be transmitted in a period, independently of variable size. Radner shows that the optimal hierarchy has two properties (see the lower left tree in Figure 2). First, they are asymmetric- in order to stay busy while those below process raw data, those above should also be involved in raw data processing. Second, and for the same reason, they involve skip-level reporting, so that even the very top manager is involved in reading and receiving messages. The reason is

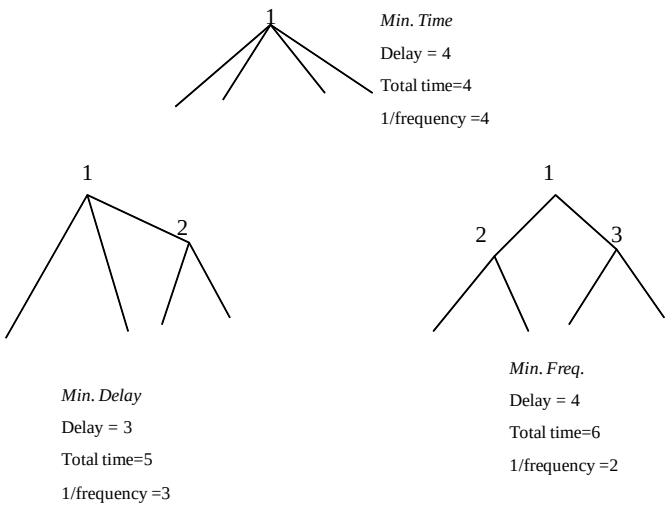

Figure 2: Optimal hierarchies to minimize total time, delay, and to maximize throughput (minimize frequency) under the assumption of Radner (1993) that processing a message takes 1 unit of time. The organization that minimizes delay is the one obtained by Radner (1993); the one that min. frequency is obtained by Bolton and Dewatripont (1994).

that this allows for using the time they would spend waiting for messages to arrive, minimizing delay.

Bolton and Dewatripont (1994) derive a more 'standard' looking hierarchy by maintaining the fixed and constant cost assumption but assuming that cohorts of data arrive all the time. Agents can improve the rate of processing output by concentrating on some problems, and thus the objective is to maximize throughput. Figure 2 illustrates these two approaches

The best applications of these ideas take place in the more recent work of van Zandt who introduces decision making models where the delay costs are endogenous rather than given by a parameter. Specifically, the model takes place in a real time information processing setting, where agents must take decisions based on the information they process. In this world, more delay means decisions are based on stale information. In van Zandt (1999, 2003), the decision problem is a resource allocation problem in which the payoff functions change over time according to some stochastic process and the resource allocation decisions must take place in each period. This work may allow organizational theorists a deeper understanding of the relationship between information, uncertainty and organizational structure. For example, it is natural in this work that delayering takes place as the business environments gets more turbulent, since increasing turbulence makes decision making based on perfect aggregation of stale information particularly useless. Indeed, as Simon (1973) noted ". . . if attention is the scarce resource, then it becomes particularly important to distinguish between problems for decisions that come with deadlines attached (real-time decisions), and problems that have relatively flexible deadlines. Rather different systems are called for to handle these different kinds of decisions."

3.2 Endogenous communication: Organizational languages

Codes, which are a shared technical language between workers, form an important part of the communication infrastructure of firms and organizations (Arrow 1974). Miscoded knowledge may lead to ambiguity, confusion, misunderstanding and inefficiency in communication and production. An optimal design of codes needs to trade off between specialization and commonality. On the one hand, a specialized code facilitates communication within a particular function that performs a task, but limits communication between functions that perform various tasks and thus makes coordination between tasks more costly. On the other hand, a broad common code improves coordination across tasks at the expense of less precise and more costly communication within tasks. In this subsection, we use a simplified variant of Crémer, Garicano and Prat (2007) to explore the effects of the attributes of tasks and the synergies between tasks on the design of codes and the interplay of optimal codes and organizational structure.

A team of two workers, worker 1 and worker 2, are employed to perform a task. As in the case of tacit knowledge and vertical communication, if worker 1 is not able to perform the task or solve the problem associated with the task, he can ask for help from worker 2. However this kind of vertical communication is limited by two forms of bounded rationality. First, both workers have a limited ability to learn codes which allow for the identification of exact problems. Second, they have a limited ability to solve problems that involve incomplete information. An example would be a team that is composed of a salesman and an engineer to serve clients, who have problems with products or services. The salesman can classify problems raised by clients but not perfectly. The engineer and the salesman have to rely on a previously specified and agreed code to transmit information coarsely. In order to make the intuition more transparent, we carry on this example to interpret the following model and use salesman for worker 1 and engineer for worker 2. The basic implications apply to many other tasks and occupations.

Task z is drawn from a distribution function F_z with the probability density function f_z on a set Ω. That is clients approach salesmen with a problem that demands a solution z with probability $f_z > 0$ from Z. The salesman, after reviewing the problem, selects a message from a *code* and transmits it to the engineer. Formally, a code C is a partition $\{\Omega_1, \Omega_2, ..., \Omega_K\}$ of the set Ω, where the subscript of the Ωs represents a word k that gives the information that the problem z belongs to the subset Ω_k. The breath of word k is n_k, the number of of events that Ω_k contains when Ω is finite or the 'size' (the Lebesgue measure) of Ω_k when Ω is a continuum.

For simplicity, we normalize the firm's profit from solving a client's problem to 1 and its target is just to minimize the expected 'miscommunication' or diagnosis cost between the

salesman and the engineer, defined as

$$D(C;F) = \sum_{k=1}^{K} p_k d(n_k), \tag{1}$$

where p_k is the frequency of a word k being sent. The per unit miscommunication cost, which can be regarded as the 'diagnosis cost' of the problem, depends on the precision of the information (the breath of word) sent by the salesman. It is natural to assume that the cost is increasing in the breath of k as less precise information brings about more costly communication. This simple specification leads to some intuitive results: a code should use precise words for frequent events and vaguer words for more unusual ones; a more unequal distribution of events increases the value of the creation of a specialized code, since the precision of the words can be more tightly linked to the characteristics of the environment.

The following two-word-code example illustrates the main ideas (the reader should refer to the paper for the analysis in the general case). Suppose that a salesman deals with problem $z \in [0,1]$ drawn from a distribution with density $f(z) = (1-b) + 2bz$,with $b \in [-1,1]$ being a measure of the evenness of the distribution. At $b = 0$, the distribution is uniform and the distribution becomes more uneven when b deviates more from 0. We also assume that codes can have at most two words, $K = 2$, and that the diagnosis cost is linear in the breath of word. Denoting with F the cumulative distribution function, the optimal code problem is:

$$\min_{z} F(z)\, z + (1 - F(z))\,(1 - z)$$

This optimization yields a unique cutoff $\widehat{z}$ that splits the set Ω into two words: $\Omega_1 = [0, \widehat{z})$ and $\Omega_2 = [\widehat{z}, 1]$. At $b = 0$, $\widehat{z} = \frac{1}{2}$. Since each problem is equally likely to occur, there is no need to use codes with different precision. When $b \neq 0$, $\widehat{z} = \frac{1}{6b}\left(3b - 2 + \sqrt{(3b^2+4)}\right)$. Obviously $f'(z) > 0$ and $\widehat{z} > \frac{1}{2}$ if $b > 0$, and $f'(z) < 0$ and $\widehat{z} < \frac{1}{2}$. A more precise code is used to deal with more frequent problems. It can be shown that $\widehat{z}$ deviates further from $\frac{1}{2}$ when $|b|$ deviates more from 0, which implies that a more specialized code is adopted when the distribution of problems is more unequal.

Integration and Separation As we have seen, optimal codes are designed to facilitate vertical communication between workers that perform the same tasks within the same organization unit. In many situations, communication is horizontal and takes place between people that perform different tasks in different working units. Then tailoring codes to the needs of particular agents in an organizational unit may be costly as it limits the set of agents among whom the codes are useful. The design of optimal codes needs to take into account the possible synergies across tasks and organization units. Two organizational units that face similar tasks will not find a common code too costly and should therefore be integrated through the same code.

We extend the simple two-word-code model discussed above to allow two services or functional units A and B. Each of them is composed of one salesman and one engineer. We focus on two possible organizational forms as shown in Figure 3 (Panel A and Panel B): (1) Separation (the two units use different codes); (2) Integration (the two units share the same code). To generate a need for coordination, there must be a potential synergy among the two services, which we model as follows. Customers arrive randomly, and there may be excessive load in one service and excessive capacity in the other. If that happens, the two services benefit from diverting some business from the overburdened service to the other. Formally, suppose that salesmen from services A and B deal with consumers from two different distributions F_A and F_B,

$$F_i(z) = (1 - b_i)\, z + b_i z^2, \qquad i = A, B$$

with $b_A = b$ and $b_B = -b$ and $b \in [-1, 1]$. Let z_i^* be the cutoff between words of each service, with (by symmetry) $z_B^* = 1 - z_A^*$, and $D_i^*(b)$ the expected diagnosis cost in either service as in (1).

Each engineer has the ability to attend to the needs of at most one client. Salesmen bring sales leads randomly to each engineer. The arrival process is as follows:

$$y = \begin{cases} 0 & \text{with probability } p, \\ 1 & \text{with probability } (1 - 2p), \\ 2 & \text{with probability } p, \end{cases}$$

where p belongs to the interval $[0, 1/2]$. This arrival process captures the effect of the variability in the expected number of clients of each type. If p is low, then each salesman is likely to find one client per period of each type. When p is high, although on average still 1 client is arriving, it is quite likely that either none or 2 will arrive. Thus p measures the importance of the synergy between the two services: a high p means that the services are likely to need to share clients, while a low p means that each service is likely to have its capacity fully utilized. [4]The profit of the firm when it solves a client's problem is 1. The per-client diagnosis costs is φ: if the engineer knows that the client's characteristics fall in an interval of size s, his diagnosis cost is $s\varphi$. We assume that the diagnosis cost is sufficiently high to ensure that information must transit through a salesman before being sent to an engineer ($\varphi > 1$) but not so high that profit risks becoming negative ($\varphi < 2$).

An *integrated organization* requires that a salesman from service unit A explain to an engineer in B the needs of his customer. Such a cross-unit explanation requires a common code in both services. It is intuitive that the common language is the one that would be chosen

[4]The restriction on φ ensures positive profits. It also ensures that information must transit through a salesman before being sent to an engineer; indeed an engineer without information on the client's problem would have diagnosis costs greater than the profits obtained from solving it.

when the density of tasks is the average of the two densities of the two services.[5] In this simple example, since both services have opposing distributions, the average problem density is uniform. The optimal code has two equally imprecise words, with each word identifying the sales lead as coming from one half of the distribution. The total profits are:[6]

$$\Pi_I(p, b, \varphi) = 2\left(1 - p(1-p)\right)\left(1 - \varphi D\left(0\right)\right).$$

In a *separated organization*, where the two services use different codes, the expected profit is:

$$\Pi_S(p, b, \varphi) = 2(1-p)(1 - \varphi D(b)).$$

The organization should be integrated rather than separated if the between service improvement in communication (measured by the synergy gain) is larger than the within service loss in precision due to the worsening of the code used:

$$\frac{1 - p(1-p)}{1-p} \geq \frac{1 - \varphi D(b)}{1 - \frac{1}{2}\varphi}.$$

The result characterizes the determinants of the trade-off between separate, well-adapted codes optimized for within-service communication, and broader common codes that allow for between-service communication. Separate codes are preferable when synergies p are relatively low, when the underlying probability distributions confronting the different units are sufficiently different(b is high), or when diagnosis costs (φ) are high so that there is a high premium on communicating precisely. As a result, increases in synergies, in the equality of the distributions or decreases in diagnosis costs increase code commonality.

Translator and Hierarchy An alternative to integration to exploit the synergy between two distinct units is to introduce a hierarchical superior as a translator, who enables services with different codes to cooperate. For instance, if salesman A has two customers, he communicates to the translator the type of the "extra" customer in the code used in service A. The translator will search for z, and then he will transmit the information to engineer B in the code used in service B. (Panel C in Figure 3).

Assume that hiring a translator requires incurring a fixed cost μ; since the translator is specialized in language, her diagnosis cost is lower than that of the engineers. The optimal organization choice depends crucially on communication costs and the translator's advantage. Hierarchies are more efficient when communication costs are high, whereas low communication costs favor their replacement by common codes and horizontal communication. Consider

[5] For a formal proof, see Corollary 1 in Cremer, Garicano and Prat (2007).

[6] The probability that a problem is solved is the sum of 1) $1 - 2p$, the probability that only one problem arrives and is passed to the engineer within the same unit; 2) p, the probability that two problems arrive and one is always passed to the engineer within the same unit; 3) p^2, the probability that two problems arrive and one is passed to the engineer in other unit that has no problem arriving.

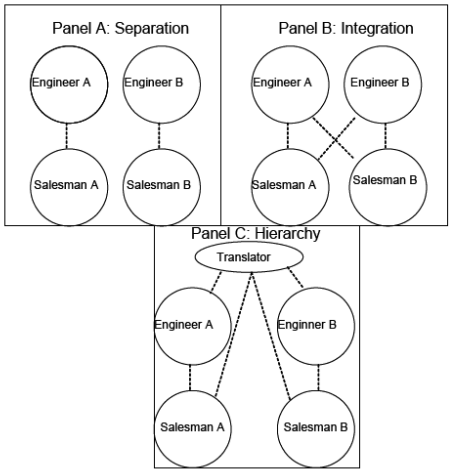

Figure 3: Communication in Three Possible Organizational Forms

first the comparison between translation and separation. Translation incurs the fixed cost μ and increases diagnosis costs, but makes inter-service communication possible and thus allows the services to profit from any existing synergies. If the diagnosis cost φ is low, the extra communication cost incurred by translation is low and the net benefit is likely to be high. Thus, translation is more likely to beat separation when φ is low. Consider the choice between translation and integration. Translation saves on communication cost by allowing services to keep efficient service-specific codes – thus translation is likely to beat integration when φ is high, since communication savings are more important when φ is high. Thus if the fixed cost μ of hiring a translator is low enough, there exists an interval of φ for which the hierarchical structure is optimal.

The theory of optimal codes helps us understand the relationship between decentralization and information technology which has been widely discussed both in the economics literature and in the business press. Accounting systems, human resource and other organizational databases are codes in Arrow's (1974) sense. In recent years, the management of these codes within firms has become more centralized, while communications have become less hierarchical and while, at the same time, decision making has become more decentralized. Robert J. Herbold, Chief Operating Officer for Microsoft from 1994 to 2001, described this apparent paradox as follows: "standardizing specific practices and centralizing certain systems also provided, perhaps surprisingly, benefits usually associated with decentralization." This paradox reflects the rationale behind the theory: better management of communication codes substitutes bureaucracies and allows for decentralization.

3.3 Endogenous knowledge acquisition

Under cognitive costs, individuals choose not to acquire all the available knowledge. In fact the decision of how much knowledge to acquire involves an expectation of both how useful this knowledge is and the extent to which other individuals may be called upon for help. In this case, organizations are useful because they allow individuals to bring to bear more knowledge, thereby leveraging the knowledge of multiple individuals in solving problems. A recent literature has tackled this issues, and thus allows us to explore the role of organization members as problem solvers, as was suggested in the March and Simon (1958, p 25) remark quoted in the introduction.

Essentially, the key advantage of organization is that it allows for an increase in the utilization rate of knowledge. While a particular piece of knowledge may be too unusual to be acquired by an individual, individuals working together will find it in their interest to do so. Of course, this is the key force for specialization in the horizontal division of labor sense, and often takes place without organization- we do, after all, have doctors and lawyers selling their wares in the market.

Organization, however, may be needed when knowledge is tacit and thus problems hard to presort, or identify ex ante. In this case, a hierarchy allows for the matching of problems with the right experts, as Garicano (2000) shows.

Following Garicano (2000), we define a random variable Z as the knowledge content of a task, which also indicates the problem that the worker will confront when performing the task. Let $\Omega \subset \mathbb{R}^+$ be the set of all possible problems and $A \subset \Omega$ be the set of problems that a worker is able to solve, referred to as his "knowledge set". When production starts, a problem $Z \in \Omega$ is drawn from an a priori known distribution F, referred to as the "knowledge distribution" of a task. The problem is solved and the task is completed if the realized knowledge content is within the worker's knowledge set, namely, $Z \in A$.

In this single task production workers acquire knowledge to solve problems. The optimal level of knowledge is determined by a comparison between the marginal expected value of additional knowledge and the marginal cost of acquiring this additional knowledge. Suppose that the cost of acquiring a knowledge set A (learning all the problems in A) is proportional to its size, i.e. the 'number' of problems in it (formally, its Lebesgue measure) $\mu(A)$. For example, $\mu(A) = z$ if $A = [0, z]$. The cost (for any worker) of acquiring knowledge set A is $c \cdot \mu(A)$, where c is a constant. The expected output x of a worker is 1 (the value of a completed project) multiplied by the probability that he has the knowledge to complete the project. Without loss of generality, we can re-order problems so that $f(.)$ is downward sloping; then the argument in f indexes the frequency of the problems. For a continuous and nonatomic F, a worker in autarchy confronting such a production function and knowledge distribution of a task maximizes expected net output y:

$$E[y] = \Pr(Z \leq z) - cz = \int_0^z f(\varphi)d\varphi - cz.$$

The optimality condition is simply[7]

$$f(z^*) = c, \tag{2}$$

which equates the marginal value of acquiring knowledge to the marginal cost: he learns those problems which are 'common enough' to justify investing in them.[8] The worker's optimal knowledge level is z^* and his knowledge set is $[0, z^*]$.

When knowledge is tacit, it is hard to formalize, express, store and transfer in some form of code. In order to solve a problem, workers need to discuss, clarify and check the information encompassed in the problem to be solved, irrespective of whether they know the answer or not. Usually this communication is between workers and managers and thus vertical. The process involves person to person and often face to face conversations and joint work. The communication cost is mainly the opportunity cost of time that could otherwise be devoted to production. Certain types of technological progress such as e-mail and video conferencing may reduce communication cost. It is also possible that the cost diminishes through frequent and repeated interactions.

Consider a simple organization with $n+1$ team members to carry out production that involves problem solving as before. There are two organizational alternatives: a one-layer structure in which all members devote their time to production and a two-layer structure with n production workers and 1 manager who can help the workers to solve problems. Suppose the workers perform the same task independently. The one-layer organization acquires knowledge to maximize their output $\max_{\{A_i\}} R_1(n) = \sum_{i=1}^{n+1}[F(A_i) - c\mu(A_i)]$.By the assumption of identical and independent distribution and the linear cost function, the optimal condition is reduced to the first order condition: $f(z_i) = c$. .

In a two-layer organization, there is a manager who may acquire more knowledge and spend time helping production workers who cannot deal with their own problem due to the limitation of their knowledge. However, help incurs communication costs: it takes time for a worker to propose a question and for a manager to figure out a solution. For simplicity, we assume that a request from a worker incurs a fixed helping cost h, which is proportional to the production time of the worker and borne only by the receiver for notational simplicity. Then the organization's target is to

[7] Assuming that learning something is worth it, that is as long as $f(0) > c$.

[8] Throughout this paper, we assume the regularity conditions for existence of optimum are satisfied. If the the density function $f(Z)$ is nonincreasing, the second order condition is always satisfied and the solution is unique.

$$\begin{aligned} \max R_2(n) &= \sum_{i=1}^{n}[F(A_i \cup A_m) - c\mu(A_i)] + t_m^p F(A_i \cup A_m) - c\mu(A_m) \qquad (3) \\ &\text{Subject to} \\ 1)\ t_m^p + t_m^h &\leq 1; \qquad 2)\ t_m^h = \sum_{i=1}^{n} h[1 - F(A_i)]t_m^p; \qquad t_m^p \geq 0 \end{aligned}$$

Here A_m is the manager's knowledge set and A_i is each worker's knowledge set. By the downward sloping assumption of $f(z)$, $A_i = [0, z_w]$ and $A_m = [0, z_m]$, where z_w and z_m are the knowledge level acquired by each worker and the manager respectively. As a result, $A_i \cup A_m = [0, z_m]$;[9] t_m^p is the manager's time devoted to production and t_m^h to helping workers.

Compared to (**??**), the two-layer organization allows for a division of labor and maybe more knowledge acquisition. The manager plays a key role in this process: she is able to leverage her knowledge – it will be worthwhile to learn unusual problems, since she can use it to answer questions from an entire team. But this advantage comes with two costs. One is the cost of acquiring additional knowledge. The other is that helping others competes away her time for production. The communication cost can be seen from the constraints. The first constraint says that the overall time for the manager is limited to a normalized unit. Since time is always valuable, this constraint will bind at optimum. The second constraint is essentially an identity that equates the communication time from both sides of the communicators (time answering questions must be equal to time asking questions)

In the optimum, $t_m^p = 0$ and $t_m^h = 1$. That is the manager completely specializes in problem solving. This is because if it pays to spend the first fraction of time leveraging time to help some workers then it is profitable to spend all other units of time in helping and not producing. Then the objective function is reduced to

$$\begin{aligned} \max R_2(z_m, z_w, n) &= nF(z_m) - cnz_w - cz_m \\ &\text{subject to} \\ 1 &= [1 - F(z_w)]hn. \end{aligned}$$

The solution is pinned down by the conditions:

$$\begin{aligned} nf(z_m) &= c; \qquad (4) \\ F(z_m) &= c[\frac{1 - F(z_w)}{f(z_w)} + z_w]; \qquad (5) \\ 1 &= nh[1 - F(z_w)] \end{aligned}$$

[9] Here we assume that the manager needs to know the worker's knowledge in order to solve the problem. The analysis applies to the case in which the knowledge sets of the manager and workers are not overlapping.

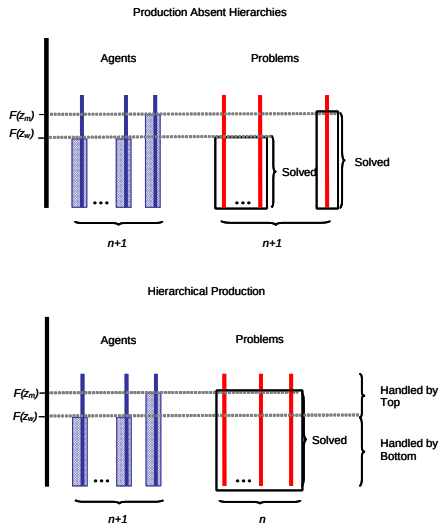

Figure 4: *The benefit of hierarchy is that it allows the manager to leverage his knowledge in problem solving ($F(z_m)$) by combining it with the time of less knowledgeable workers, so that the team solves more problems; the cost is that the number of problems tackled is lower than in autarchy ($n+1>n$), since 1 unit of time (the manager's time endowment) is spent in communication.*

From the first line in (4), the optimal knowledge level in the two-layer organization is higher than in the one-layer organization: the marginal value of manager knowledge is larger, as it is spread over n workers. This is exactly the effect of knowledge leverage which allows for specialization and a higher knowledge level. Second, given that $f(z_m)$ is decreasing in z_m, a more knowledgeable manager attains a larger span of control. Third, the number of workers increase in the knowledge they acquire since a more knowledgeable worker asks fewer questions and gives more time to other workers. A comparison of the hierarchical production and the production absent of hierarchy is illustrated in Figure 3.

A full model without restrictions on the number of layers is developed in Garicano (2000). In this model, a knowledge hierarchy efficiently integrates tacit knowledge. Members in the organization specialize either in production or in solving problems and only one class specializes in production (referred to as production workers). Those who specialize in problem solving are managers allocated at the higher level of the hierarchy. Production workers learn to solve the most common problems; managers or problem solvers learn the exceptions. The higher is a member in the hierarchy, the more unusual the problems she is able to solve. Moreover the organization has a pyramidal structure, each layer possessing a smaller size than the previous one.

Knowledge hierarchies allow more knowledgeable workers to specialize in exceptional problems. This "management by exception" was well stated by Alfred Sloan (1924, P. 195), who in describing his job, claimed that "we do not do much routine work with details. They never get up to us. I work fairly hard, but it is on exceptions..., not on routine or petty details." In

the presence of communication costs, knowledge chains or hierarchies emerge with the more knowledgeable placed on the top as managers. These managers acquire knowledge about exceptional problems and specialize in solving problems from their subordinates. A knowledge hierarchical structure is advantageous only if the size of organization is large enough– leveraging the knowledge of highly skilled managers (where knowledge can be broadly construed as knowledge of opportunities, clients etc.) requires assigning them better workers so that they can be protected from the 'dumb' questions anyone else could deal with.

The model yields implications about the interplay between organizational change and the improvements in information and communication technologies (ICT). Unlike the usual treatment of ICT as homogeneous, it allows us to distinguish two types of progress in ICT. One type is related to cheaper acquisition of knowledge (reductions in c), resulting, for example, from the introduction of Enterprise Resource Planning (ERP). The other is related to more efficient communication or helping in the model (reductions in h), resulting, for example, from improvements in IP-based and wireless communications. Decreases in the cost of both communicating and acquiring knowledge increase the level of organizational knowledge and in general lead to an expansion of organization. However, they have opposite impacts on the discretion of the production workers (bottom of the knowledge hierarchy or chain) and the managers (at the upper positions of the knowledge hierarchy or chain). Cheaper acquisition of knowledge increases the knowledge scope of production workers and thus reduces the frequency of interventions from above. On the other hand, better communication of knowledge reduces the knowledge scope of the production workers and increases the need for interventions. This challenges the view that improvements in ICT lead to more delegation of power and flattened organization. Bloom et al (2009) find, using detailed international plant-level data and ICT information, that the evidence is consistent with the theory that we have described.

4 Bringing Together Incentives and Cognition

A recent stream of research in organizational economics attempts to combine incentive theory and team theory. We focus on two new developments: a set of papers that introduces incentives in the classical team-theoretic coordination problem and a literature that studies costly communication under moral hazard.

The multi-tasking literature first observed the existence of a trade-off between coordination and motivation and argued that if incentives are endogenous one may expect that low-powered incentives may be optimal (Holmstrom and Milgrom (1991, 1994), Holmstrom (1999)). This literature does not actually model the coordination tasks. Closer in spirit to the approach of the previous section in trying to marry the study of coordination with the analysis of the incentive problems is the work of Hart and Moore (2006), who study how to allocate authority over the use of assets when agents with several assets (coordinators)

can have ideas involving the common use of several of these assets, and when agents are motivated by their own interest rather than that of the organization, and of Hart and Holmstrom (forthcoming), who show that whereas independent firms coordinate their activities too little, integrated firms have a tendency to realize too many synergies, neglecting private benefits of mangers and workers. Thus organizational structure clearly affects incentives. Athey and Roberts (2001) show that the allocation of authority affects the trade-off between giving agents incentives for decisions and for effort provision when only a broad signal that adds both incentives and the output from the project is available.

A more recent literature (Alonso, Dessein, and Matouschek 2008, Rantakari 2010) establishes an explicit link with the coordination models of the bounded rationality approach. Essentially, like in team theoretical models in the past, information is distributed but unlike there, it can be communicated, at a cost. Specifically, the limit to communication is that managers are biased and thus not truthful.

Adopting Alonso, Dessein, and Matouschek's (2008) notation, there are two production units, each of which has the profit function:

$$\pi_i = K_i - (d_i - \theta_i)^2 - \delta(d_i - d_j)^2, i, j = 1, 2$$

where θ_i are random variables and the $d_i's$ are decisions. This is a classical quadratic team payoff function: the first term, the distance between the θ_i and d_i, measures the adaptation of unit 1 to its environment; while the second term, the distance between the decisions of both units, measures the coordination among them; δ is the value of coordination.

The modelling innovation lies in the introduction of private incentives. Managers $1, 2$ are assumed to privately observe their own θ_i; moreover they are (exogenously) biased towards the profits of their own units. In particular, manager 1 cares about a convex combination of his own and $2's$ profits, $\lambda\pi_1 + (1 - \lambda)\pi_2$, where λ is the bias. Given this bias, it is straightforward to introduce communication costs: managers can communicate their θ_i, but their messages m_1, m_2 are cheap talk, that is they are unverifiable. As a result, only messages which are incentive compatible can be transmitted

With these elements in place, we can now study the effect of organizational structure. In a decentralized structure, decisions are given, for example for manager 1, by:

$$\max_{d_1} E[\lambda\pi_1 + (1 - \lambda)\pi_2|\theta_1, m_1, m_2].$$

Alternatively, both managers can send messages to a "central" manager, who is assumed to be unbiased but uninformed a priori. In this case, he will solve:

$$\max_{d_1, d_2} E[\pi_1 + \pi_2|m_1, m_2].$$

Given this set up, one can now analyze the impact of coordination on organizational

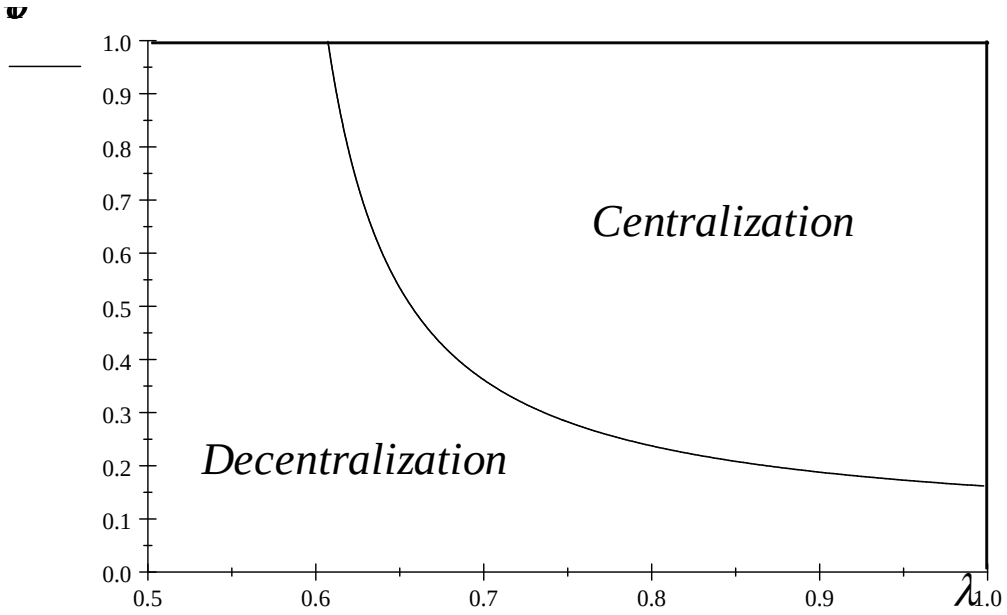

Figure 5: Incentives vs. Coordination: Optimal organizational form in Alonso, Dessein, Matoushek (2008) as a function of the value of coordination δ and of the incentive misalignment λ.

structure. The main insight is that decentralization can be preferred even if coordination is important. The reason is that when managers individually care about coordination or synergies, they have larger incentives to be truthful to each other. This increases truthtelling, and as long as managerial bias is not too large, decentralization can still be preferred.

Dessein, Garicano and Gertner (2010) follow a similar approach, but allow for endogenous decision making biases. That is, the organization can choose, by determining the incentive formula, how much a manager is biased in favor of his own unit. Managers remain optimally biased, since effort provision still requires that managers be compensated for their own results; as a result, managers are too narrow minded, decision making is too uncoordinated, and synergies hard to capture.

A second research area where development is occurring is the analysis of the interplay between organizational structure, individual incentives and communication costs. Endogenous communication between parties with different objectives has been a central topic in microeconomics (see Sobel (2010) for a comprehensive survey). If organization is a way to overcome communication limits, then it is important to understand how the cognitive limits of individual agents interact with the incentives they face in determining the amount of information that is transmitted in equilibrium.

As Dewatripont and Tirole (2005) note, economic theory has focused on two polar cases of communication: hard information that can be acquired costlessly and soft information

that can never be verified. In many circumstances, the truth lies in between. While the statement of a theorem is cheap talk in se, I can verify its accuracy if I am willing to check its proof carefully, but that requires a certain mathematical knowledge and a time investment. However, I often choose to believe the theorem based on information about its authors' reputation and incentives. Dewatripont and Tirole consider both modes of communication: issue-relevant information (writing and reading the proof) and transmission of "cues" about the sender's credibility (their academic reputation).

The amount of communication that occurs in equilibrium depends on the parties' incentives. There is a strategic complementarity between the communication effort of the sender and that of the receiver. An exogenous decrease in the stake of one of the two parties lowers the equilibrium effort of the other party.

An important organizational lesson is that, when the two parties' interests become more aligned, issue-relevant communication gives way to soft information transmission. This creates a linkage between incentives and communication mode. Optimal organizations must foster informal communication among groups of agents with aligned interests, but must also realize that communication among other agents may need costly transmission of hard information. The formality of the communication channel is a decreasing function of the two parties interest congruence.

Calvo, de Marti and Prat (2009) extend Dewatripont and Tirole (2005) beyond two agents and embed it in the classical team-theoretic environment used by organizational economics (with the addition of incentives). As in Dessein and Santos (2006), every agent i faces a normally-distributed local source of uncertainty x_i. Before choosing their actions, agents can communicate with their colleagues. Agent i receives from agent j a message

$$y_{ij} = x_j + \varepsilon_{ij}^s + \varepsilon_{ij}^r,$$

where ε_{ij}^s is a normally distributed variable with precision (1/variance) s_{ij}, which represents the noise under the control of sender j, and ε_{ij}^r is another normally distributed variable with precision r_{ij}, which represents the noise due to the receiver i. Precise communication is costly on both ends. The sender j pays cost $k_s^2 s_{ij}$ to achieve precision s_{ij}. The receiver i pays $k_r^2 r_{ij}$ to achieve precision r_{ij}, where k_s^2 and k_r^2 are parameters.

The payoff function of each agent is quadratic:

$$\omega\left(x, a_1, \ldots, a_n\right) = -\left(d_{ii}\left(a_i - x_i\right)^2 + \sum_{j \neq i} d_{ij}\left(a_i - a_j\right)^2 + k_r^2 \sum_{j \neq i} r_{ji} + k_s^2 \sum_{j \neq i} s_{ij}\right)$$

where the terms of the form d_{ii} measure the importance of adaptation while the terms of the form d_{ij} capture the need for agent j to coordinate with agent i, which may differ from the need for i to coordinate with j.

In this two-stage model, first all agents select communication intensities s_{ij} and r_{ij}, then,

after observing signal realizations, they choose actions a_i. The model admits a linear solution:

$$a_i^* = b_{ii}x_i + \sum_{j \neq i} b_{ij}y_{ij}$$

The main technical contribution is to find a close-form expression for the *influence coefficients* b_{ij}, which measure how much the action a_i selected by agent i is influenced by the signal y_{ij} sent by j, as well as one for the communication intensities s_{ij} and r_{ij}. For any organizational objective (a matrix of interaction coefficients d_{ij}), one can determine the equilibrium influence structure and the equilibrium communication structure.

As we shall see in the next section, predictions on s_{ij} and r_{ij} are testable by using data on electronic communication within organizations.

For instance, the result can be used to analyze equilibrium communication in a matrix form organization. For concreteness, interpret agents as units of a multinational firm. Suppose that every i is associated to two attributes: function f_i and country c_i. There are n_f functions and n_c countries, so there are a total of $n_f n_c$ units. Assume that

$$d_{ij} = \begin{cases} G & \text{if } f_i = f_j \text{ and } c_i = c_j \\ F & \text{if } f_i = f_j \text{ and } c_i \neq c_j \\ C & \text{if } f_i \neq f_j \text{ and } c_i = c_j \\ L & \text{otherwise} \end{cases}$$

Communication will be more intense on dimensions where interactions are stronger. So, if $C > F$, we will see more communication within a country than within a function. More importantly, the model can be used to see what happens when units are grouped into divisions. In Crémer (1980), if two units belonged to the same division they would communicate for free. Here, instead, grouping has no direct effect on communication technology: it only means that those units share the same payoff, namely they become part of the same team.

If, for instance, we group country functions together, we obtain two effects. As expected, communication intensity between units that belong to the same country increases. More interestingly, communication between units that belong to the same function decreases. If units belong to the same division, they have a stronger incentive to coordinate well with each other. This increases their loss if they let individual members coordinate with members of other divisions. Hence, they have less incentive to invest in communication. This captures an important organizational decision: whenever we offer a collective incentive to a set of agents, we foster team work between them but we also push them apart from the rest of the organization. In other words, there exists a trade-off between the cohesion of a specific division and its integration within the firm at large.

5 The foundations and implications of bounded rationality

In this final section we identify some recent developments – ranging from purely theory to data availability – that we see as crucial for the research agenda in the future. First, researchers have tried to get better microfoundations for cognitive costs. Second, recent research aims to go beyond understanding organizations, by building organizations into models of labor markets. Richer views of firms, if useful, should contribute to better understanding of markets. Finally, we study how the recent flood of data internal to organizations, including email, personnel data etc. allows for direct testing of theories that, up to know, where hard to test.

5.1 Modeling Cognitive Costs

If cognitive limits lie at the core of models of organization, it is important that the way we model such limits is consistent, general and well-documented. Since Alan Turing's work, computer science has developed ways to think about various limits to information processing. The key concept is that of complexity class. Problems are categorized depending on the amount of resources necessary to solve them, given the best known algorithms. This means that the complexity notion depends on the type of resource that we are focusing on.

The most widespread notion is that of computational complexity (e.g. Papadimitriou 1994). The resource under consideration is computation time, which is proportional to the number of basic operations that a Turing machine should perform in order to solve the problem. A literature has developed at the intersection of economics and computer science to study mechanism design from a computational complexity viewpoint. This area, often referred to as algorithmic mechanism design (for a survey see Nisan, Roughgarden, Tardos, and Vazirani 2007, Chapter 2), is quite active but it has tended to focus on a different set of problems, mostly related to auctions and competitive allocation mechanisms.

While computational complexity is central to computer science, it has not yet had an impact on organizational economics. One possible reason for the lack of intellectual arbitrage is that computational complexity focuses mainly on finding exact solutions – or excellent approximations – to difficult but well-defined problems, like the traveling salesman's problem. Unfortunately, this approach does not appear to capture, even at a stylized level, the sort of challenges that firms face, and have therefore found virtually no application in organizational economics.

An alternative definition of complexity has found wider application in issues of interest to organizational economists. Instead of analyzing computation time, we now focus on the communication burden. How much information must organization members transmit to each other given a certain organization structure? This question is closely related to Hayek's view that society faces the enormous task of aggregating "dispersed bits of incomplete and often contradictory knowledge which all separate individuals possess." In Yao's (1979) definition of communication complexity, there are multiple agents, each of whom possesses some informa-

tion. The agents must solve a problem (establish the truth of a proposition). Jointly, they have enough information to do it but individually they do not. A communication protocol is a description of a dynamic exchange of information among the agents. The communication burden of a protocol is the maximum number of bits that must be exchanged by agents (in the worst possible instance within a certain class of problems). The communication complexity of a certain class of problems is given by the protocol with the lowest communication burden. In a Hayekian view, we are looking for ways to optimize the process of aggregating dispersed knowledge.[10]

Segal (1995) uses communication complexity to understand coordination by authority. Consider two agents who must agree on a joint action from a finite set A. Each agent i observes a signal from the set $\Theta_{i.}$. There is no conflict of interest between the agents. They both want to select the optimal action given their joint information. However, the agents incur a communication cost that is proportional to the number of bits they transmit to each other. A communication protocol is an extensive form game that describes the dynamic communication occurring between the two agents (which may depend on the values of signals being communicated). An end game represents a joint action given the history of signals. The communication complexity of a protocol is the worst-case number of bits transmitted.

One particular communication protocol is coordination by authority. The only information exchange that occurs consists of one agent indicating an action to the other agent. Clearly, the communication burden is of the order of $|A|$. Coordination by authority saves on communication but can lead to a suboptimal action, whenever information is dispersed between the two agents. Protocols where the agents also 'talk' about their information can lead to better decisions.

However, Segal proves a negative result. Let the size of a problem be $n = |A|$. A protocol is of polynomial complexity if, for some γ and k, its communication burden can be bounded above by γn^k for all n. Segal identifies a class of problems such that the additional expected payoff generated by moving from coordination by authority to any protocol of polynomial complexity tends to 0 as $n \to \infty$.

In the class of problems used to prove this result, a joint action is 'good' if both agents receive a positive signal about it. If at least one of the agents receives a negative signal, the action is bad. If one assumes that signals are independent and that the proportion (but not the expected number) of good actions goes to zero as $n \to \infty$, then finding a good action requires a great deal of communication. When an agent describes an action for which he received a positive signal, the probability that that particular action is also positive for the other agent goes to zero. That is true for any finite set of actions. Hence, to do better than coordination by authority, agents must be prepared to describe an infinite number of actions. The theorem implies that asymptotically the only way to improve on coordination by authority is to use protocols that are exponential in the size of the problem n, making the

[10] This notion is related to mechanism design with bounds on communication (surveyed in Marschak 2006).

communication burden potentially unsustainable.

To understand the economic intuition behind the theorem, it is important to keep in mind two underlying assumptions. First, the class of problems used to prove the negative result relies on agents' signals being independent. The presence of strongly interrelated signals could make communication less time consuming. Second, each agent has an ordering of the set of actions A, but this ordering is not common. If the agents had a common dictionary, communication would be less costly. Instead of describing an actions, agents could just indicate its position in the dictionary. Loosely speaking, the theorem applies to situations where agents have limited prior information about the environment they face and they do not share a common language to communicate effectively. The result can then be interpreted as saying that, to engage in useful communication beyond coordination by authority, the organization must possess either useful prior information on the optimal solution or a shared organizational language – as in Arrow (1974).[11]

A third source of cognitive limits is memory. Economists have explored settings where agents have bounded recall (see for instance Dow 1991, Piccione and Rubinstein 1997, Wilson 2004). Miller and Rozen (2010) analyze moral hazard in teams where remembering is costly: agents can allocate their time between production and monitoring. The authors characterize the optimal organizational arrangement in the presence of these cognitive constraints. Unlike in traditional settings, agents, especially those with highly uncertain tasks, make empty promises – namely they do not fulfill commitments to undertake certain tasks and they are not punished.

A related form of bounded rationality relates to categorization, which can be measured in terms of the number of states that an automaton needs to implement a certain procedure. This notion has found application in repeated games (Rubinstein 1986) and more recently as a way of assessing the cognitive requirements of different choice rules (Salant 2010). A choice procedure yields a benefit in terms of payoff from the alternative chosen and a cost in terms of state complexity. The optimal procedure may display dynamic framing effects, such history dependency and a recency effect.

Finally, as mentioned above, a fundamental feature of the set of problems that firms face is that they are ill-defined. As March and Simon (1958, p 190) noted: "Because of the limits of human intellective capacities in comparison with the complexities of the problems that individual and organizations face, rational behavior calls for simplified models that capture the main features of a problem without capturing all its complexities." Firms use simplified and potentially incorrect representations of the environment in which they operate. As they realize that there may be a discrepancy between model and reality, they look for

[11] Communication complexity has also been used to study the hold-up problem (Segal 1999). The kind of contracts that can prevent inefficient outcomes due to the ex-post renegotiation (Maskin and Moore 1999) require a large communication burden as the number of possible contingencies increases. In a complex environment, the parties' inability to foresee all possible trades ex ante combined with the cost of describing them ex post makes it difficult to eliminate the hold-up problem.

organizational solutions that tolerate a certain degree of model misspecification.

Maradasz and Prat (2010) explore an example of organizational response to model uncertainty. They re-visit the classical screening problem under the assumption that the firm operates on the basis of a simplified model of the true distribution of types. The authors show that mechanisms that are optimal when the firm knows the true distribution can perform very poorly when the firm uses an approximate type space, even as the model converges to the truth. Instead, the authors identify a class of mechanisms that yield a near-optimal payoff even if when they are based on an approximate type space. In Simon's terminology, these mechanisms can be seen as satisficing rather than optimal: they achieve an outcome that is adequate even if the model is slightly misspecified.

This is not an exhaustive list of the potential sources of cognitive limits that organizations face. Economists have examined the effects of other forms of bounded rationality in other contexts. In particular, there is a growing literature on the response of firms to boundedly rational consumers (see Spiegler 2011 for a comprehensive survey). There is great potential for transposing some of those ideas from consumers to employees. For instance, how should an organization be structured if its members have time inconsistent preferences or biased beliefs?

5.2 Putting Firms into Labor Markets

An area where the organization models we have reviewed have great potential is the study of the interdependence of labor market and firm structure. It is clear that changes in wages and wage inequality are a function of the internal restructuring of firms: as the division of tasks between managers and workers change, the returns to skill change. Similarly, the optimal organization of firms responds to changes in wage schedule, the extent and cost of offshoring and other market equilibrium phenomena. Embedding optimally organized firms inside markets is a challenge for which the theories we have reviewed are well suited.

Specifically, if organizations are devices that aim to leverage the knowledge of multiple individuals to solve problems (as in Garicano, 2000), then changes in how individuals communicate and how costly they find it to solve problems will affect not only how these individuals are organized, but also the return to their skills and thus their wages. That is, wages are not just affected by human capital and productivity, but by the coordination and communication costs among individuals.

For example, consider the problem of understanding how labor markets and the organization of firms react to information technology changes. Following the theory in Garicano (2000) discussed in Section 3.3 above, Bloom et al (2009) conduct an empirical study of how changes in communication (networks etc.) and information access (e.g. access to databases) change the structure of firms (decentralized decisions at different levels, spans of control, etc.). Consistent with the theory, they find that cheaper information access decentralizes– as more information is available, people become generalists and need less help; while cheaper commu-

nication centralizes, as people specialize more and need more help, they rely more on experts, on stars, and on knowledge and information located at corporate headquarters. This clearly has implications for the labor market, which are analyzed by Garicano and Rossi-Hansberg (2006) (GRH).

GRH develop a competitive model of the labor market with workers with heterogeneous cognitive ability. Recall from the treatment in 3.3 that, given a density of problems confronted $f(.)$, a worker with knowledge z can solve a fraction $F(z)$ of problems; that he needs help with probability $1-F(z)$; that helping him costs $h(1-F(z))$ where h is the cost of communication or helping; and thus that a manager who is matched with workers with knowledge z can help $n = 1/(h(1-(F(z))$ workers. GRH assume that the helping or communication cost h is equal for all workers, but that different workers can learn how to solve problems at a different cost– specifically, smarter workers are those who incur a lower cost of learning to learn the same interval of problems. They specify the cost of acquiring knowledge a as a function of skill α and technology , so that the cost of acquiring knowledge z is $a(\alpha;t)z$ and comparative (and absolute) advantage holds: high ability types have a comparative advantage in knowledge acquisition. This allows for a study of the impact of communication and information acquisition cost on wages, inequality and organization. The problem is solved in two stages: for a given set of wages and assignment, the organization of the firm (given by knowledge acquisition, spans of control and layers) must be optimal. Then the equilibrium in the labor market is obtained, in which (1) Agents choose occupations to maximize utility; (2) firms choose the skill of their employees, their knowledge, and their number; (3) firms make zero profits and (4) labor markets clear, that is, the matching of workers to managers is such that supply and demand are equalized at every point of the skill distribution.

Among the findings of the analysis, GRH show that when information technology reduces the cost of acquiring information, individuals gain more autonomy, the number of layers of organization decrease, and wage inequality increases primarily within occupational classes. In this world being a bit more skilled makes workers a bit better off. The implications of better communication technology are different. Those who use this technology to leverage their knowledge, managers and experts, are better off, as they can leverage it more; while those who rely on others acquire less knowledge as a result and are worse off. Communication technology thus results in an increase in wage inequality between occupational classes.

A similar, although slightly simpler, framework can be used to study the implications of offshoring. The key feature of offshoring is that it allows the formation of cross-country trades- it thus allows for matches across different countries. To study the impact of offshoring on the labor market and the organization of firms, Antras, Garicano and Rossi-Hansberg (2006) propose a model where the distribution of problem solving knowledge in the population z is exogenously given; the distribution of skills in the population is given by a cumulative distribution function $G(z)$, with density $g(z)$. Equilibrium is similar to GRH, except that now knowledge acquisition is exogenous.

From this analysis, Antras, Garicano and Rossi-Hansberg (2006) analyze the impact of the formation of cross-country teams between North and South on wages and organization, where the "North" has more skills and thus can specialize in problem solving and the "South" specializes in production . For example, they conclude that when technology permits the pairing of low skilled workers from the South with high skilled workers from the North, within-worker inequality increases in the South as a result of changes in matching: globalization improves the quality of the managers with whom southern workers are matched, thus raising the productivity of these workers, and thereby leading to an increase in their marginal return to skill. This effect is reinforced by an occupational choice effect: more agents become workers, hence increasing the range of abilities in the worker skill distribution.

They also show that organizational forces imply that the effect in the North is less clear. On the one hand, we have a traditional labor market effect: low skill workers in the North face increased competition from southern workers and this tends to reduce their marginal return to skill. On the other hand, organization plays a key role: when more low skill agents are available, the time of high skill managers becomes more scarce, and workers who are better able to economize on this time become relatively more valuable. As a result, the value of more skilled workers relative to less skilled ones increases, as does the difference between the ability of the managers they are matched with. When either communication costs or the skill overlap are sufficiently low, so that high skill managers are particularly valuable and scarce, this last effect dominates and globalization increases wage inequality not only in the South but also in the North.

5.3 Interpreting Activity Data

The cognitive costs discussed in this survey have a strong temporal dimension. Communicating or processing information takes time. If we can observe how agents allocate their time (*activity data*), we can make inferences on the cognitive costs they incur. This means that organizational theories based on cognitive costs have a potentially large empirical relevance. For instance, in Geanakoplos and Milgrom (1991), managers choose how to allocate their limited time τ to different tasks. Their model makes a rich set of predictions on time allocation, compensation, and performance – all variables that are potentially testable.

Traditionally, activity data was hard to obtain. An ethnographic study was needed to record how workers spent their time. The presence of an outside observer was both costly an intrusive. However, the IT revolution has made the collection of activity data much simpler. How organization members spend their time can be gleaned from email data, calendaring software, social networks, etc.

Bandiera, Guiso, Prat and Sadun (2009) provide an example of how a simple model with cognitive costs can be used to guide the analysis of activity data. They observe the activities of 103 CEOs of top-600 Italian firms during one randomly selected week. In particular, they observe time devoted to communication – meetings, phone calls, events, etc... – which

occupies the vast majority of the CEOs' work time. Inspired by Geanakoplos and Milgrom (1991), they hypothesize that the CEO faces n activities and allocates non-negative time vector $(x_1, ..., x_I)$ to the activities.

The firm's production function is

$$Y = \sum_{i=1}^{n} \alpha_i x_i$$

The vector α describes the value of the top manager's time in all possible activities and it is determined by the firm technology and environment.

The CEO can also produce some personal rent (e.g. networking), with production function

$$R = \sum_{i=1}^{n} \rho_i x_i$$

The vector ρ depends on characteristics of the CEO and the institutional and economic environment he operates in.

The total cost of time for the CEO is $C = \frac{1}{2}\sum_{i=1}^{n} x_i^2$. There is an increasing marginal cost of devoting time to one particular activity, due either to the onset of boredom or to a lower time-efficiency. The CEO's payoff is

$$u = bY + (1-b)\,R - C.$$

The parameter b plays an important role. It denotes the alignment between the firm's interests and the CEO's – implicit or explicit – incentive structure. If $b = 1$, the firm and the CEO have perfectly aligned interests. If $b = 0$, the CEO only pursues personal interest.

It is easy to see that the optimal time allocation satisfies

$$\frac{\hat{x}_i}{\hat{x}_j} = \frac{b\alpha_i + (1-b)\,\rho_i}{b\alpha_j + (1-b)\,\rho_j}.$$

In the extreme case of perfect alignment ($b = 1$), the CEO devotes time to activities in proportion to the relative value of the activities to the firm:

$$\frac{\hat{x}_i}{\hat{x}_j} = \frac{\alpha_i}{\alpha_j}.$$

More generally, the relative allocation of time across activities will be determined both by the firm's needs and by the CEO's preferences. To put some structure on the problem, assume that activities can be grouped into two sets: I_Y and I_R. The first set – let's call elements of I_Y productive activities – contains activities that benefits the firm but not the CEO ($\alpha_i > 0$ and $\rho_i = 0$), while the second one contains activities – networking activities – that are only beneficial to the CEO ($\alpha_i = 0$ and $\rho_i > 0$). This leads to three sets of testable

implications:

1. In equilibrium, the cross-sectional correlation between the time $\hat{x}_i$ that the CEO devotes to a particular activity and the total time the CEO spends at work is positive if and only if the activity is productive.

2. In equilibrium, the cross-sectional correlation between the time $\hat{x}_i$ that the CEO devotes to an activity i and firm's productivity $\hat{Y}$ is positive if and only if activity i is productive.

3. The governance measure $\hat{b}$ is positively correlated with time spent on an activity if and only that activity is productive.

Bandiera, Guiso, Prat and Sadun (2009) combine the activity data discussed above as well as standard data on firm's performance and governance. Such information, given the three predictions above, can be used to understand the relative productivity of different activities. In particular, there is a debate over whether CEOs devote time to activities outside the firm in the interest of the company or for their own benefit (Malmendier and Tate 2009).

If the time that CEOs devote to communication is split between time spent only with employees of the firm and time spent also with outsiders (e.g. consultants, investors, politicians, etc.), one finds that the share of time a CEO spends with outsiders (insiders) is: (1) negatively (positively) correlated with time spent at work; (2) uncorrelated (positively correlated) with firm performance; (2) negatively (positively) correlated with quality of governance. In light of the three predictions above, the evidence indicates that spending time with outsiders is at the margin less productive than spending time with insiders, and that CEOs do it in part for their own benefit, especially if they work for a firm with poor governance. This interpretation is consistent with the need for CEOs who operate in environments with an uncertain incentive structure to maintain high visibility in the business community in order to generate new employment opportunities (Khurana 2002).

Palacios-Huerta and Prat (2010) use email data to analyze communication within firms. The underlying idea, put forward by Arrow (1974) and formalized by Calvo, de Marti and Prat (2009), is that intra-firm communication patterns are endogenous and should reflect the priorities of the organization. Agents should allocate more time to writing email to agents that they consider more "important". Hence, the relative importance of agents can potentially be captured through email traffic measures.

A natural candidate for such a measure is an eigenvector index such as that used in Google's PageRank or in the more sophisticated bibliographical impact factor measures. An email from A to B is treated like a A citing B's work. By a fixed-point argument, the importance of B is then given by the sum of emails received weighted by the importance of the senders, which in turn is given by the weighted importance of email received, etc. Such a measure – also called the Invariant Method – is the only one with a number of desirable properties (Palacios-Huerta and Volij 2004).

Of course, there is no guarantee that this email-based measure will work in practice. So, Palacios-Huerta and Prat use a database of email traffic between all top executives in a large Spanish retail company to determine how this index correlated with actual organizational outcomes. The impact factor of an executive – computed uniquely on the basis of email data – turns out to be strongly correlated with: (i) the executive's rank in the corporation (the agent with the highest factor is the CEO and nowhere in the company a subordinate has a higher factor than his boss); (ii) the executive's salary (controlling for rank); (iii) the chance that he will be promoted or dismissed (a positive/negative deviation from the impact factor predicted by the executive's rank and salary is predictive of that executive being promoted/dismissed in the future).

6 Conclusions

In the past three decades, organizational economists have almost entirely focused on how organizations solve incentive problems. Although cognitive limits where sometimes assumed (as in the incomplete contracts literature) to justify some contracting shortcuts, there has been limited work on what the relevant cognitive limits are and what organizational response we should expect.

This survey has reviewed the literature on how organizations respond to the cognitive limits of their members, starting from early contributions especially team theory and ending with the recent resurgence of interest in this area. We hope to have convinced the reader that combining cognition and incentives is both a possible and extremely promising line of research.

We have also identified three external stimuli that are affecting the economic study of organizations. First, computer science and other disciplines are providing powerful and coherent ways to model cognitive costs. Second, other disciplines within economics – especially labor economics, industrial organization, and international trade – feel an increasing need to enrich and refine their theoretical predictions by opening up the 'black box' of organizations. Third, the availability of activity data will make it easier for economists to quantify what happens within firms and to test competing organizational theories.

References

[1] Ricardo Alonso, Wouter Dessein, and Niko Matouschek. "When Does Coordination Require Centralization?" *American Economic Review*, 98(1), 145-179, March 2008.

[2] Arrow, Kenneth. J. 1974. The Limits of Organization. New York, NY: Norton.

[3] Bandiera, Oriana, Luigi Guiso, Andrea Prat, and Raffaella Sadun. What Do CEOs Do? Working Paper, London School of Economics, 2010.

[4] Bloom, Nick., Luis Garicano., Raffaella Sadun, and John Van Reenen. 2009 "The distinct effects of information technology and communication technology on firm organization." CEP Discussion Papers, 927. Centre for Economic Performance, London School of Economics and Political Science.

[5] Bloom, Nick, and John Van Reenen. 2007. "Measuring and explaining management practices across firms and countries." *Quarterly Journal of Economics*, 122 (4). pp. 1351-1408.

[6] Bolton, Patrick and Matthias Dewatripont, 1994. "The firm as a communication network." Quarterly Journal of Economics. Vol. 109,4, pp. 809-839

[7] Brynjolfsson, Eric, and Lorin Hitt. 1996. "Paradox Lost? Firm-level Evidence on the Returns to Information Systems Spending." Management Science 42(4), 541-58.

[8] Calvo-Armengol, Antoni, Joan de Marti, and Andrea Prat. 2009. Communication and Influence in Organizations. Working paper, London School of Economics.

[9] Crémer, Jacques. 1980. "A Partial Theory of the Optimal Organization of a Bureaucracy." *Bell Journal of Economics* 11 (Autumn): 683–93.

[10] Crémer, Jacques. 1993. "Corporate Culture: Cognitive Aspects." *Industrial and Corporate Change*, 3(2): 351–386.

[11] Crémer, Jacques, Luis Garicano., and Andrea Prat. 2007. Language and the Theory of the Firm. *Quarterly Journal of Economics* 122, 373-407.

[12] Dessein, Wouter, Robert Gertner, and Luis Garicano. "Organizing for Synergies." Forthcoming, *American Economic Journal-Micro.*

[13] Dessein,Wouter and Tano Santos. "Adaptive Organizations." *Journal of Political Economy*, 114(5): 956–995, 2006.

[14] Mathias Dewatripont and Jean Tirole. Modes of Communication. *Journal of Political Economy* 113(6): 1217–1238, December 2005.

[15] Dow, James. Search Decisions with Limited Memory. *Review of Economic Studies* 58(1): 1–14, 1991.

[16] Garicano, Luis. 2000. "Hierarchy and the Organization of Knowledge in Production;". *Journal of Political Economy.*

[17] Garicano, Luis. and Esteban Rossi-Hansberg. 2006. "Organization and Inequality in a Knowledge Economy", Quarterly Journal of Economics, 121(4): 1383-1435.

[18] Geanakoplos, John and Paul Milgrom, 1991. "A theory of hierarchies based on limited managerial attention," Journal of the Japanese and International Economy.

[19] Grant, R. M. 1996. "Toward a Knowedge-based Theory of the Firm," Strategic Management Journal. 17 (December), 109-122.

[20] Groves, Theodore. "Incentives in Teams, Econometrica, Vol. 41, No. 4 (Jul., 1973): 617-631.

[21] Hayek, Friedrich A. von. 1945. "The Use of Knowledge in Society." American Economic Review. Vol. 35: 519-30.

[22] Holmostrom, Bengt R. 1977. "On incentives and control in organizations", Doctoral Dissertation, Stanford University.

[23] Holmstrom, B. 1999. The Firm as a Subeconomy. Journal of Law, Economics and Organization.

[24] Hurwicz, Leonid, 1973. "The Design of Mechanisms for Resource Allocation", American Economic Review, vol 63, No. 2:. 1-30.

[25] Ichniowski, C., K. Shaw., and G. Prennushi. 1997. "The Effects of Human Resource Management Practices on Productivity." American Economic Review. 87(3): 291-313.

[26] Khurana, Rakesh. *Searching for a Corporate Savior: The Irrational Quest for Charismatic CEOs*. Princeton, NJ: Princeton University Press. 2002.

[27] Kogut, B., and U. Zander. 1992. "Knowledge of the Firm, Combinative Capabilities, and the Replication of Technology." Organization Science, 3, 383-397.

[28] Jacob Marschak and Roy Radner. *The Economic Theory of Teams*. Yale University Press, 1972.

[29] Madarasz, Kristof and Andrea Prat. Screening with an Approximate Type Space. CEPR Discussion Paper 7900, 2010.

[30] Malmendier, U, and G. Tate, "Superstar CEOs"Quarterly Journal of Economics, November 2009, vol. 124(4): 1593-1638.

[31] Marschak, T. Organization Structure. In Hendershott, T, *Handbook in Information Systems*, Volume I: 201–284, Elsevier, 2006.

[32] Milgrom, P., and J. Roberts. 1990. "The Economics of Modern Manufacturing: Technology, Strategy, and Organization." American Economic Review. June, 511-528.

[33] Milgrom, P., and J. Roberts. 1992. Economics, Organization and Management. Englewood Cliffs, NJ: Prentice Hall.

[34] Milgrom, P., and J. Roberts. 1995. "Complementarities and Fit Strategy, Structure, and Organizational Change in Manufacturing." Journal of Accounting and Economics.

[35] Miller, David A. and Kareen Rozen. Monitoring with Collective Memory: Forgiveness for Optimally Empty Promises. Cowles Foundation Discussion Paper 1698. April 2010.

[36] Nelson, R., and S. Winter. 1982. An Evolutionary Theory of Economic Change. Harvard University Press, Cambridge, MA.

[37] Nonaka, I., and H. Taceuchi. 1995. The Knowledge-creating Company: How Japanese Companies Create the Dynamics of Innovation. Oxford University Press, New York.

[38] Palacios-Huerta, Ignacio and Andrea Prat. Measuring the Impact Factor of Agents within an Organization Using Communication Patterns. CEPR Discussion Paper 8040, 2010.

[39] Ignacio Palacios-Huerta and Oscar Volij. The Measurement of Intellectual Influence. *Econometrica* 72(3), 963-977, May 2004.

[40] Piccione, Michele and Ariel Rubinstein. On the Interpretation of Decision Problems with Imperfect Recall. Games and Economic Behavior 20: 3–24, 1997.

[41] Prat, Andrea. "Should a Team Be Homogeneous?" *European Economic Review* (46)7: 1187-1207, 2002.

[42] Prescott, E. C., and M. Visscher. 1980. "Organizational Capital." Journal of Political Economy. 88(3): 366-82.

[43] Yingyi Qian, Gerard Roland, and Chenggang Xu. "Coordination and Experimentation in M-Form and U-Form Organizations," *Journal of Political Economy*, 114(2), pp.366-402, 2006.

[44] Radner, Roy (1993), The organization of Decentralized Information Profcessing. Econometrica, 62, 11090 - 1146.

[45] Rantakari, Heikki. "Governing Adaptation." *Review of Economic Studies*, Vol. 75, Issue 4, pp. 1257-1285, October 2008.

[46] Roberts, J. 2004. The Modern Firm. Oxford University Press.

[47] Rosen, Sherwin, 1983. "Specialization and Human Capital" Journal of Labor Economics, 1(1), 43-49

[48] Salant, Yves, 2010. "Procedural analysis of choice rules with applications to bounded rationality", forthcoming, American Economic Review.

[49] Sah, R., and J. Stiglitz. 1986. The Architecture of Economic Systems: Hierarchies and Polyarchies," The American Economic Review. LXXVI, 716 – 727.

[50] Simon, Herbert. 1953 “A Formal Theory Model of the Employment Relationship.” Econometrica 19: 293-305.

[51] Simon, Herbert A. 1973. “Applying Information Technology to Organization Design” *Public Administration Review*, Vol. 33, No. 3 (May - Jun., 1973): 268-278.

[52] Simon, H. A. 1991. "Bounded Rationality and Organizational Learning." Organization Science, ", 125-134.

[53] Sloan, Alfred. 1924. "The Most Important Thing I Ever Learned about Management." System 46 (August).

[54] Sobel, Joel. Giving and Receiving Advice. Working paper, UCSD. August 2010.

[55] Spiegler, Rani. *Bounded Rationality and Industrial Organization.* Oxford University Press, 2010.

[56] Teece, D. J., G. Pisano and A. Shuen. 1997. "Dynamic Capabilities and Strategic Management," Strategic Management Journal. 18(7): 509-533.

[57] Van Zandt, Timothy (1999). “Real Time Decentralized Information Processing as a Model of Organizations with Boundedly Rational Agents” Review of Economics Studies, 66, 633-658.

[58] Van Zandt, Timothy (2003), “Real Time Resource Allocation with Quadratic Payoffs”, *Mimeo.*

[59] Wilson, Andrea. Bounded Memory and Biases in Information Processing. Working Paper, Princeton University, 2002.