Decentralized Information Processing in the Theory of Organizations*

Timothy Van Zandt†
Princeton University

Published 1999 in
Contemporary Economic Issues, Vol. 4: Economic Design and Behavior,
edited by Murat Sertel. London: MacMillan Press Ltd. Chapter 7, pages 125-160.
(This is a prepublication version.)

Abstract

Bounded rationality has been an important theme throughout the history of the theory of organizations, because it explains the sharing of information processing tasks and the existence of administrative staffs that coordinate large organizations. This article broadly surveys the theories of organizations that model such bounded rationality and decentralized information processing.

Author's address:
Department of Economics
Princeton University
Princeton, NJ 08544-1021
USA

Voice: (609) 258-4050
Fax: (609) 258-6419
Email: tvz@princeton.edu
Web: www.princeton.edu/~tvz

*Presented at the 1995 World Congress of the International Economic Association in Tunis.
†This research was supported in part by grant SBR-9223917 from the National Science Foundation.

Contents

1. Bounded rationality and organizations

Although models with bounded rationality—defined broadly to mean those in which agent behavior departs from the paradigm of effortless full rationality—are appearing in every field of economics, only in the theory of organizations has bounded rationality been an important theme throughout the history of the field. There is a reason for this difference. The classical model of rational choice has been a good approximation and powerful tool in studying consumer and producer theory, markets and price determination, imperfect competition, trading in financial markets, and most other topics in economics. Only recently have fields matured enough that some outstanding open questions require more accurate models of human decision making. In the theory of organizations, on the other hand, the rational model leads to uninteresting models of organizations, in which one manager or entrepreneur can run a firm or economy of arbitrary size. Without bounds on information processing capacity, it is impossible to explain the sharing of information processing tasks that is such an important part of the interaction between members of organizations, and to explain the existence and functioning of the administrative apparatus that are such important components of organizations (as documented, for example, by Chandler (1966, 1990)).

The purpose of this article is to give a broad and historical overview of the economic theory of such joint or *decentralized* information processing in organizations. This topic is inherently interdisciplinary, with contributions not only by economists but also by researchers in management and information systems, computer science and operations research. However, this already broad survey focuses primarily on the treatment of this topic by economists.

Section 2 covers two methodological issues. Section 3 reviews early debates on the computational efficiency of planning bureaus versus markets and of firms versus markets, in which the theme of decentralization of information processing was important. Section 4 examines decompositions of decision problems. Formal modeling of information processing constraints is taken up in Section 5, which reviews models that emphasize communication constraints, and in Section 6, which reviews models that emphasize computation constraints. Section 7 concludes by describing a few directions for future research. An appendix mentions some related research in other fields.

2. Two methodological issues

2.1. Information processing constraints and constrained optimality

Most of the literature on information processing in organizations has adopted what I call the *constrained-optimal* approach to modeling bounded rationality.

This means, most importantly, that the modeler explicitly incorporates information processing constraints or costs into a computation, communication or decision problem. A decision procedure then specifies not only a decision rule but also the communication or computation procedure by which the decision rule is computed. The second defining is that the modeler characterizes the feasible decision procedures that are optimal—or that at least are good by some criterion, if it is too difficult to characterize optimal procedures—according to a performance criterion that may reflect computation or com-

munication costs.

Another common approach to modeling bounded rationality is to characterize selected non-optimal decision rules in a model without explicit information processing constraints. The non-optimal decision rules are informally motivated by complexity considerations. This approach has been used most extensively in the non-rational learning literature. It is not always clear which non-optimal decision rules are the most relevant, but this literature has many results on the asymptotic properties of non-optimal learning rules or evolutionary mechanisms that are robust with respect to changes in the specification of the rules or mechanisms.

These approaches are not pure substitutes. It is possible to define a set of feasible decision procedures that incorporates information processing constraints, and then characterize selected non-optimal procedures in order to incorporate further unmodeled bounds on rationality. For example, one can construct learning or evolutionary models in which the constraints on the daily processing of information are modeled explicitly and the properties of non-optimal learning rules or evolutionary mechanisms that adjust the decision procedures are studied. An example of this in game theory is Binmore and Samuelson (1992), and two examples in the theory of information processing in organizations are Jordan (1995) and Miller (1996).

However, there are reasons for instead characterizing constrained-optimal procedures when the model incorporates information processing constraints:

1. If the goal is organization design, then constrained-optimality is the appropriate objective.
2. Economic theory is not meant simply to provide accurate predictions, but also to isolate and elucidate the relationships between phenomena. A characterization of constrained-optimal decision procedures isolates the effects of the modeled information processing *constraints*, which can restrict the set of feasible decision rules in interesting and complex ways.
3. Constrained-optimality can be a good approximation in a descriptive theory, at least compared to picking arbitrary suboptimal procedures. This does not presume that the agents whose bounded rationality is modeled, such as the information processing agents in an organization, can effortlessly choose constrained-optimal decision procedures. Instead, it presumes that the selection of the alternatives or the design of organizations takes place on a time scale that is different from that of the daily information processing tasks that are in the model. For example, the decision procedures and organization structures observed in modern firms are the result of many years of incremental learning, imitation and adaptation, and may become quite good in environments that are stationary, even though the human hardware on which these procedures run is slow enough that the daily cost of and delay in implementing the procedures is significant.

2.2. Communication constraints and bounded rationality

The bounded rationality of economic agents means that there are limits on their ability to communicate (i.e., to formulate and send messages and to read and interpret messages) and to calculate with information in the brain. Both communication and computation constraints appear implicitly or explicitly in the papers reviewed here.

There are also constraints on the physical transmission of messages, which I view

as less significant than the human communication constraints. (E.g., compare the cost of copying and mailing a research paper to 100 people with the cost of 100 people reading the research paper.) These two types of constraints have quite different effects on decision making procedures. There is no fundamental constraint on the bandwidth (capacity) of information transmission to or from any one agent or on information acquisition by any one agent. For example, if a principal needs to communicate with n agents, he can establish n telephone lines or postal routes so that in one period he can send a message to or receive a message from each of n agents, however large n is. The capacity of a single person to read and write messages, on the other hand, is fundamentally limited. If it takes the principal 1 period to read 1 message, then it takes him n periods to read a message from each of n agents. Hence, whether 10 agents report to one principal or 5 of the agents report to one principal and the other 5 report to a second principal does not affect the information transmission costs, but does affect the time it takes for all the messages to be read.

However, sometimes a model that is meant to reflect human communication costs, but that allows individuals to compute any function of the information they have read, is isomorphic to a model with unboundedly rational decision makers in which either it is costly to acquire or to transmit information. Modeling bounded rationality in this way still offers new insights, because it may suggest information acquisition or transmission costs where normally there would be none. For example, an economist modeling costly information acquisition without bounded rationality in mind would assume that, once information is acquired by a firm, it is available or can be made available to all members of the firm at no cost. When modeling bounded rationality, on the other hand, we might instead assume that when information is acquired by or generated internally by a firm, each member of the firm who uses this information must invest some time to do so.

3. Some early debates

3.1. Markets versus planning bureaus

The term *organization* is often interpreted as a tightly coordinated bureaucracy in which individuals have specific functions and common purpose. However, more broadly an organization is any collection of individuals or groups whose actions are coordinated and for which there is some criterion for evaluating the collective outcome, even if it is a weak criterion such as Pareto efficiency. Hence, markets and bureaucratic structures such as firms are alternate forms of organizations. I shall use this broad definition of organizations, and shall use the term *bureaucracy* to refer to tightly coordinated organizations such as firms.

Having thus included market mechanisms in the theory of organization, I can say that the first visible discussions in economics about alternate forms of organizations and their relative efficiency for processing information were debates about socialism from about 1910 to 1940.

Some early visions of planning under socialism held that economic activity would be directed entirely by a central authority, without any role for prices. However, it was later suggested that prices were necessarily part of efficient resource allocation mechanisms, because they arise naturally in such a constrained optimization problem.

This is essentially the argument of both Barone (1935, originally published in 1908) and Mises (1951, originally published in 1922), although the former was proposing a price-based planning mechanism and the latter was claiming that price-based planning was not possible because of the lack of private ownership and exchange of the means of production. (See Hayek (1935) for details on the origin of these ideas.)

This point was soon accepted by many economists, along with the additional point that the computation of the solutions to the planning problem, even using prices, was too large a task to be done centrally by a planning bureau and would require the communication of too much information to the planning bureau. (Hayek (1940, pp. 125-126) summarizes this consensus). Therefore, later stages of the debate argued about whether socialist economies could use competitive (decentralized) price mechanisms to allocate resources. Taylor (1929), Lange (1936, 1937), and Dickinson (1939) and others proposed iterative, decentralized price mechanisms in which the adjustment of prices was controlled by the planning bureau. Hayek (1940) and others contended that such mechanisms would be too cumbersome or slow, but these authors never presented a model of how private property markets reach equilibrium, and hence the computational efficiency of the proposed planning mechanisms and of markets in capitalist economies could not be compared. This is an important gap that is yet to be filled.

3.2. Markets versus firms

In the 1930s, while some economists were comparing the efficiency of socialism and private-property market economies, other economists, such as Kaldor (1934), Robinson (1934), and Coase (1937), were drawing attention to the fact that even in private-property market economies, many transactions take place inside firms and are not regulated by price mechanisms. They suggested that the boundaries of firms are determined, in part, by the relative efficiency of markets and bureaucracies for processing information. However, this issue was also not resolved because the authors discussed the process of managing a firm but did not make a direct comparison with how markets perform similar tasks.

One of the themes of that brief literature, which has consistently reappeared in the theory of organizations, is that information processing constraints may be a limit to firm size. It was first observed that with centralized information processing, meaning that a single entrepreneur processed all information and made all decisions, there would be decreasing returns to scale because of the fixed information processing capacity of firms. On the other hand, it was also noted that as firms grow, more managers are hired and information processing is decentralized. Kaldor (1934, p. 68) responded that full decentralization of the coordination task is still not possible:

> You cannot increase the supply of co-ordinating ability available to an enterprise alongside an increase in the supply of other factors, as it is the essence of co-ordination that every single decision should be made on a comparison with all the other decisions made or likely to be made; it must therefore pass through a single brain.

This suggests that information processing constraints will lead not only to decentralized information processing, but also to decentralized decision making, in which multiple agents make decisions based on differing information.

These themes have been further developed by Williamson (1975, 1985) and others in the field of transaction economics, and in the formal models of information processing described below.

4. Decomposition

The debates in the 1930's on socialism and on the management of firms, reviewed in the previous section, both observed that decentralizing information processing and decision making economizes on communication costs and distributes the information processing tasks. Such decentralization requires that decision problems be *decomposed* into tasks that can be assigned to different agents. The decomposition can be into detailed steps such as adding two numbers or transmitting one bit of data, but it then takes some work to build up a model with recognizable patterns of information processing and decision making. In contrast, it can be easy to interpret a less detailed decomposition of a decision problem into recognizable subproblems. The study of such decompositions was prevalent in the theory of organizations from the mid 1950's to the early 1970's, aided by advances in mathematical optimization methods.

Consider the following classical decision problem. An organization has n units, and its decision problem is

$$\begin{aligned} \max_{\langle x_1,y_1,\ldots,x_n,y_n\rangle\subset\mathbb{R}^{2n}} \quad & \textstyle\sum_{i=1}^n \pi_i(x_i,y_i) \\ \text{subject to:} \quad & \textstyle\sum_{i=1}^n x_i \le x\ . \end{aligned} \tag{1}$$

For example, each unit is a division in a firm, $\langle x_1,\ldots,x_n\rangle$ is an allocation of a quantity x of a common resource such as capital, y_i is a local decision variable for unit i, and π_i is the profit function of unit i. The data in this decision problem are the profit functions $\{\pi_i\}_{i=1}^n$ and the total available resource x.

For $i = 1,\ldots,n$, define $f_i\colon \mathbb{R}\to\mathbb{R}^2$ by the following profit maximization problem for unit i given a price or shadow price:

$$f_i(p) \;=\; \arg\max_{\langle x_i,y_i\rangle\in\mathbb{R}^2} \pi_i(x_i,y_i) - px_i\ . \tag{2}$$

Define $f_0\colon \mathbb{R}^{n+1}\to\mathbb{R}$ by the following Walrasian price adjustment process:

$$f_0(p,x_1,\ldots,x_n) \;=\; p + \alpha\left(\textstyle\sum_{i=1}^n x_i - x\right) \tag{3}$$

for some $\alpha > 0$. Define a dynamic iterative process by selecting an initial shadow price $p^0\in\mathbb{R}$ and setting, for $t\ge 0$, $\langle x_i^t, y_i^t\rangle = f_i(p^t)$ and $p^{t+1} = f_0(p^t, x_1^t,\ldots,x_n^t)$. Under certain assumptions, $\langle x_1^*, y_1^*,\ldots,x_n^*,y_n^*\rangle$ is a solution to (1) and p^* is the associated shadow price if and only if $\langle p^*, x_1^*, y_1^*,\ldots,x_n^*,y_n^*\rangle$ is a steady state of the dynamic process. Under additional assumptions, the process converges to a steady state. This then defines an algorithm for calculating an approximate solution to (1), or an exact solution if the process converges in finitely many steps.

This process is said to be a *decomposition* of (1), not because it is iterative and makes use of shadow prices, but because the problem is divided into subproblems (the maximization problem (2) defining f_i and the formula (3) defining f_0) that are smaller than the original problem and that do not use all the data of the original problem, except

through the indices that are communicated in each iteration and that coordinate the subproblems.

The decomposition can be viewed as a decentralized decision procedure, as follows. There is a resource manager who calculates f_0 in each iteration. For each $i = 1, \ldots, n$, there is a unit manager who calculates f_i. In each iteration, the resource manager communicates p^t to the unit managers, and each unit manager i responds by communicating x^t to the resource manager. One advantage of such decentralization is that the computation tasks are shared among the $n + 1$ managers. If the unit managers are endogenously endowed with the information about their profit functions, then another advantage of this decentralized procedure is that these managers do not have to communicate their entire profit functions to the resource manager or some central office.

Two pioneers of research on decomposed decision procedures were Kenneth Arrow and Leonid Hurwicz, whose early work includes Arrow et al. (1958), Arrow and Hurwicz (1960), and Hurwicz (1960).[1] Their motivation was twofold. One was to provide foundations for economic planning, both economy-wide and within firms. Is it possible, as proposed by Taylor (1929), Lange (1936, 1937), and Dickinson (1939), to establish practical planning procedures that calculate the desired social outcome without having to gather all information about the economy in a central office? The other motivation was to provide foundations for competitive equilibrium in markets. As a static equilibrium condition, competitive markets are informationally decentralized in that it is possible to check whether a given allocation and prices are an equilibrium by inquiring about each agent's net demand for the prices, rather than having to learn each agent's entire utility or profit function. Can markets reach, i.e., calculate, the equilibrium with similar decentralization, as suggested by Walras (1954) and Pareto (1927)? Arrow et al. (1958) and Arrow and Hurwicz (1960) provided (qualified) affirmative answers to these questions by devising decomposed iterative gradient methods for solving non-linear constrained optimization, such as the one outlined above.

For linear programs, a decomposed algorithm was developed by Dantzig and Wolfe (1960). Their motivation was practical. By breaking up a linear programming problem into smaller subproblems, it became possible to solve large problems for which the simplex method would exceed the memory capacity of computers. However, the application of such decompositions to descriptive and normative models of human organizations was immediately recognized as well. Dantzig and Wolfe (1960, p. 101) mention that their decomposition method for linear programs, "besides holding promise for the efficient computation of large-scale systems, …yields a certain rationale for the 'decentralized decision process' in the theory of the firm." Baumol and Fabian (1964) elaborated on this interpretation of the Dantzig-Wolfe decomposition.

In subsequent research, the planning literature studied the design of decomposed decision procedures for solving economy-wide planning problems. Heal (1973) is a good exposition of this research; Chapter 3 provides a concise outline of the issues. See Heal (1986) for a more recent survey. Furthermore, there is an enormous body of research in operations research and management science on the design of decomposed decision procedures for firms. This includes the literatures on multi-level systems (e.g., Dirickx and Jennergren (1979), Burton and Obel (1984), Van de Panne (1991), and Sethi and Zhang (1994)), aggregation/disaggregation methods (see Rogers et al. (1991) for

1. See Arrow and Hurwicz (1977) for a collection of papers on the topic by these authors; the preface contains interesting historical notes.

a survey), and abstract systems theory (Mesarovic and Takahara (1989)). Some of the research described in later sections of this survey is also based on decompositions of decision problems.

5. Communication constraints

The decomposition of decision problems shows the possibility of decentralization, but does not in itself quantify the benefits of doing so. For this, we need to model the cost of or constraints on communication and information processing. In this section, I review models of *communication* complexity, and in Section 6 I review models of *computation* complexity.

5.1. Overview

The economics literature on communication complexity includes the Hurwicz iterative procedures (message-space) literature, reviewed in Section 5.3, and team theory, reviewed in Section 5.4, both of which were initiated in the late 1950's. Starting in the 1980's, there has been some work on mechanism design that integrates communication complexity and incentive constraints; this topic is treated in Section 5.5.

The study of communication complexity in computer science dates back to the information theory of Shannon (1948). However, this paradigm considers only the complexity of a one-way flow of information, rather than the interactive communication that arises in economic models. Computer scientists started studying such interactive communication complexity in the late 1970's, with the seminal papers of Abselon (1980) and Yao (1979). I will not review this literature in any detail, but I do draw on it the following subsection when I give an overview of communication procedures and costs. Kushilevitz and Nisan (1997) is a good introduction to the many tools and results that have been developed.

Although computer science research on interactive communication complexity is motivated by problems of communication in distributed computer networks and within VLSI chips or parts of a computer, the framework is abstract and similar to that used by economists. On the other hand, the computer science literature focuses on problems in which variables and messages are drawn from finite sets, whereas the majority of economics models are based on continuous parameters. Perhaps this is why, with the recent exception of Segal (1996), economists have not made much use of this literature.

5.2. Communication procedures

The communication complexity models can all be more or less squeezed into the following framework. The scenario is an n-agent situation in which each agent i observes information $e_i \in E_i$. e_i is also called i's environment or type, and $e \in E \equiv \prod_{i=1}^{n} E_i$ is called the environment or the type profile. The problem is to choose an outcome $a \in A$, dependent on the realization of the environment, via a *communication procedure*, and thereby compute a goal function $F : E \to A$. The communication procedure specifies

the rules for exchanging information and selecting outcomes.[2]

A communication procedure could be part of a model in which the computation and incentive constraints of agents are also modeled. However, here there are no such constraints and instead there is only some measure of communication costs for each communication procedure. Given such a measure, one can conduct the following exercises: Fix a goal function and characterize the lowest-cost procedure that computes it (the usual approach in computer science), or fix only a performance measure on goal functions (e.g., expected payoff) and characterize the procedure that is optimal taking into account both the communication costs and the performance of the computed goal function.

When $n \geq 3$, we can distinguish between *broadcast* communication, in which all messages are received by all agents, and *targeted* or *networked* communication, in which messages can be sent to a designated set of recipients, upon which the cost of the message depends. We begin with broadcast communication because it is simpler.

In a broadcast communication procedure, at each moment in time all agents have the same information about the history of messages (although messages can be sent simultaneously). Therefore, if the outcome depends only on the final history of messages, then the realized outcome is common knowledge among the agents and can be physically implemented by any of them or even by an outside observer. Such a procedure is called *non-parametric*. In a *parametric* procedure, the action space is decomposed into one or more decision variables or actions after the exchange of messages, and each of these is selected privately by one of the agents. Using parametric procedures, when allowed, reduces communication costs because the outcome can depend on information that is never announced publicly.

Parametric and non-parametric broadcast communication procedures can be represented by the following:

1. An extensive game form for the n agents, with perfect information and simultaneous moves, as defined in Osborne and Rubinstein (1994, Chapter 6).[3] The moves or actions are called messages. Let H_i be the histories after which agent i sends a message and let M_{ih} be the *message space* or set of messages available to i after history $h \in H_i$. Let Z be the set of terminal histories.
2. In the non-parametric case, an outcome function $g\colon Z \to A$. In the parametric case, for each $h \in Z$, a decomposition $A_{1h} \times \cdots \times A_{nh} = A$ and outcome functions $\{g_{ih}\colon E_i \to A_{ih}\}_{i=1}^{n}$.
3. A communication rule $f_{ih}\colon E_i \to M_{ih}$ for each agent i and each $h \in H_i$.[4]

Each environment $e \in E$ determines the messages sent after each history and hence the terminal history that is reached, and the outcome $F(e)$ that is obtained after this terminal

2. The term "procedure" comes mainly from the planning literature. Hurwicz (1960, 1972) uses the terms "process" and "mechanism", and the computer science literature uses the term "protocol". I have not used "process" because the "to process", "processors", "information processing" and "stochastic processes" are also used in this paper. I have not used "mechanism" because in the mechanism design literature with incentive constraints initiated by Hurwicz (1972), the term "mechanism" eventually came to refer only to components 1 and 2 of a non-parametric procedure, as defined below. I have not used "protocol" because this term is not standard for economists.

3. Equivalently, with multiple stages and observed actions, as defined in Fudenberg and Tirole (1991, Chapter 3).

4. The communication rules and outcome functions (in the parametric case) for agent i constitute i's decision rule, and are analogous to a strategy for i in a Bayesian game.

history. $F : E \to A$ is the goal function computed by the procedure.[5]

A measure of communication costs needs to specify the cost of each message—which then determines the cost of each terminal history—and a way to aggregate the costs of the possible terminal histories. The standard way to measure the cost of a message with message space M_{ih} is by the "size" of M_{ih}, as follows.

Suppose first that all message spaces are finite; this is the approach taken by Yao (1979) and most of the communication complexity literature in computer science. If a message space M_{ih} contains k elements, then the message can be encoded with $\log_2 k$ bits, and so this is the size of the message.[6]

Suppose instead that all message spaces are Euclidean; this is true in the majority of economics models. The cost of a message from M_{ih} might be measured by $\dim(M_{ih})$. This measure is harder to interpret and there is the possibility of encoding an arbitrary amount of information into a message ("information smuggling"). Therefore, Mount and Reiter (1974, 1977) and Abselon (1980) develop smoothness restrictions on the communication rules so that several real numbers cannot be encoded into a single real number. Given such a Euclidean procedure, we can construct a discrete procedure in which each real number is mapped continuously into $(0, 1)$ and approximated by the first k bits of its binary expansion. The communication cost of a message from M_{ih} is then $k \dim(M_{ih})$. As $k \to \infty$, the goal function computed by the discrete procedure converges pointwise to the one computed by the Euclidean procedure.

Having assigned a size or cost to each message, we can sum the cost of a sequence of messages to obtain the communication cost of any history. There are several ways to then assign an overall communication cost to a procedure. The worst-case communication cost is the supremum of the costs of all possible histories. An alternative is to calculate the expected communication cost of the histories with respect to some distribution on E. For finite procedures, we can instead use $\log_2$ of the number of terminal histories; this lies between the minimum cost and maximum cost of the histories. Yet another alternative is to restrict each agent to use the same message space M_i at all times and measure the communication cost be the size of $M_1 \times \cdots \times M_n$, i.e., of the procedure's "language". Because a long message can be sent as a sequence of short messages with smaller message spaces, this measure is only meaningful if restrictions are imposed on the memory of the communication rules and outcome functions, e.g., if these can only depend on the last message sent by each agent.

Networked communication procedures can be represented by an extensive game form with incomplete information. Only the parametric case makes much sense, unless there is an outside observing who hears all messages. The selection of actions by the agents should be incorporated into the extensive game form in order to describe the information available to each agent when choosing actions. The measurement of communication costs is messy. First, some moves in the extensive game form represent private actions rather than messages. Second, costs depend on the recipients of each

5. An extension is where communication rules may be random (mixed strategies) and F must be evaluated correctly with a certain probability.

6. Information theory provides refinements on this measure using data compression that depends on the statistical distribution of the possible messages and on randomization that I have not introduced. If there are no simultaneous moves, then any finite procedure can be converted to a binary procedure in which each message space has two elements. (A long message is replaced by a sequence of one-bit messages.) As explained in Kushilevitz and Nisan (1997), it is then possible to incorporate such encoding into the extensive game form and communication rules.

message, as determined by the extensive game form.

5.3. Iterative procedures and message-space complexity

Hurwicz (1960) developed abstract *iterative* communication procedures.[7] An iterative procedure specifies for agent i a message space M_i and a response function $f_i\colon M \times E_i \to M_i$, where $M \equiv \prod_{i=1}^n M_i$. Agents simultaneously broadcast messages at each iteration. If at step t, messages $m^t = \langle m_1^t, \ldots, m_n^t\rangle$ are exchanged, then at step $t+1$, agent i sends the message $m_i^{t+1} = f_i(m^t, e_i)$. An exogenous initial message profile $m^0 \in M$ starts the process. In the non-parametric case, there is an outcome function $g\colon M \to A$ such that if the procedure is stopped after T iterations then the outcome is $g(m^T)$.

A message $m = \langle m_1, \ldots, m_n\rangle$ is stationary if $m_i = f_i(m, e_i)$ for all i. The presumption is that the procedure is run until a stationary message is reached or is approximately reached. Suppose that, for every environment e, the sequence of messages converges to a stationary message $\hat{\mu}(e)$ either asymptotically or after finitely many iterations. Then the goal function that is computed by the procedure is $g \circ \hat{\mu}\colon E \to A$.

Observe that the decomposition of the resource allocation problem in Section 4 can easily be represented by such an iterative communication procedure, as can many planning procedures. Another example is the following representation of the competitive price mechanism for classical exchange economies. The environment e_i is i's endowment and preferences, and a is an allocation. The goal is to find a Pareto optimal allocation. In period t, agent 1 sends a price p^t and each agent $i = 2, \ldots, n$ sends a net trade z_i^t. The price p^t is the gradient of agent 1's utility function when her net trade is $-\sum_{i=2}^n z_i^{t-1}$, that is, when she balances the market given the net trades of the other agents in the previous exchange of messages. Agent's i's net trade z_i^t is chosen to maximize i's utility given the Walrasian budget set for the price p^{t-1}. Observe that if the procedure is stopped at period T, the net trade of agent $i \geq 2$ is z_i^T and the net trade of agent 1 is $-\sum_{i=2}^n z_i^T$. If the procedure is stopped at a stationary message, then the resulting allocation is a Walrasian equilibrium. This is similar to a Walrasian tatônnement process or a Lange-Lehrer planning procedure, but the adjustment of prices is performed by agent 1 rather than by an auctioneer or central planning bureau.

These procedures illustrate the advantage of interactive communication over one-shot communication. To compute, with a single exchange of information, an approximate solution to the resource allocation problem would require that each agent transmits an approximation of his entire utility function. In contrast, in the iterative procedures, detailed information is eventually transmitted about the utility functions in a neighborhood of the approximate solution, but little information is transmitted about the functions in the rest of their domains.

As explained in Section 5.2, a measure of the amount of communication that takes place in each iteration is $\log_2 \#M$ if M is finite, or $\dim(M)$ if M is Euclidean. However, measuring the total communication requirements of such iterative procedures is tricky, as it depends on how many iterations are made before the calculation is stopped. If the sequence of messages only converges asymptotically to a stationary message, then total communication would have to be measured by the number of iterations needed to calculate an approximate solution to within a given error.

Rather than attempt such difficult measurement of total communication, this litera-

7. See Reiter (1977) for an introduction to the subject, and Hurwicz (1986) for an extensive exposition.

ture measures communication requirements by the amount of communication that must take place in each iteration, i.e., by the size of the message space M. This measure is meaningful, as explained in Section 5.2, because the definition of an iterative procedure requires that each round of messages depends only on the previous round of messages and on the individual environments. Thus, $\dim(M)$ is also a proxy for the complexity of the decision rules, in the sense that it measures the size of their domains.

A further simplification was made, initially at least, by ignoring the question of whether the procedures actually converge to stationary messages. That is, the function or correspondence computed by a procedure is defined to be $g \circ \mu$, where $\mu(e)$ is the set of stationary messages when the environment is e. The message-space literature thus posed the following question: Given a goal function $F : E \to A$ to compute, what is the iterative communication procedure with the smallest message space such that $g \circ \mu = F$? Since we can at least check in one round of information exchange whether a candidate message profile m^* is stationary given the environment, this exercise is similar to determining the complexity of checking whether a given answer is correct, rather than of computing an answer from scratch.

In the competitive price mechanism outlined above, the set of stationary messages is the set of Walrasian equilibria. Hence, this procedure is presumed to compute the Walrasian correspondence, even though substantial restrictions on the set of possible utility functions (environments) would be needed to ensure that such a tatônnement procedure actually converges. In the procedure, agent 1 sends normalized prices from an $(\ell-1)$-dimensional space, where ℓ is the number of commodities. Because of Walras' Law, agents $2, \ldots, n$ only need to send excess demands for $\ell - 1$ commodities. The total dimension of the message space is thus $n(\ell - 1)$. Mount and Reiter (1974) and Hurwicz (1977) have shown that, for classical economies, this competitive price mechanism is informationally efficient for achieving Pareto-optimal allocations.

In later work, researchers added the restriction that the procedures converge globally or locally, i.e., that the procedures be globally or locally stable. This requirement typically increases the minimum size of the message space. See, for example, Mount and Reiter (1987), Jordan (1987), and Reiter and Simon (1992). Also, Moore et al. (1996) considers rules for stopping in finite time.

A variation of this paradigm that is suitable for studying organizational structure has recently been developed by Marschak and Reichelstein (1995, 1996). Communication is networked, one counts the size of each communication link connecting two agents, and there are potential limits on individual communication complexity. These authors use the model to study the structure of firms for a fixed number of agents, but, as with other models with individual communication constraints, it could incorporate an endogenous number of information-processing agents.

5.4. Team theory

Another approach to modeling communication costs in organizations is team theory,[8] introduced and developed by Marschak (1955), Radner (1962), and Marschak and Radner (1972). (See Marschak (1986, Section 3) for a survey.) The team-theory and the message-space literatures are quite different. First, team theory imposes statistical

8. The term "team theory" is sometimes used more generally for any model of group decisions with no conflicts of interest, in which case it includes just about all the literature surveyed in this paper.

assumptions on the environments and includes a state-dependent payoff function so that it is possible to compare the expected payoffs of different outcome functions; these can thus be endogenous. Second, team theory studies procedures with finite depth (finitely many rounds of communication), rather than studying stationary messages of iterative procedures. Third, team theory did not develop and has not made use of a standard specification of communication costs; hence, most team theory models have studied a small number of game forms that are motived informally by communication complexity.

These three differences are interrelated. The message-space literature has studied how to compute an outcome that is optimal conditional on the pooled information of the agents, without actually communicating all the private information. In many cases, however, this has not been possible via a finite exchange of finite-dimensional messages. In contrast, by incorporating statistical decision theory, team theory can compare procedures that exchange too little information to calculate the optimum, so that ultimately there is truly decentralized decision making, in which different decision variables are controlled by different agents and decisions are not the same as those that would be taken if all information were pooled.

On the other hand, the incorporation of statistical decision theory made the measurement of communication complexity more difficult in team theory. The smoothness restrictions that make the dimension of Euclidean message spaces a meaningful measure of complexity in the message-space literature lose their bite in a statistical model. For example, let $E \equiv \mathbb{R}^2$ be the sample space of two normal random variables $\tilde{e}_1$ and $\tilde{e}_2$, and let $M \equiv \mathbb{R}$ be a message space. There is no smooth map $f\colon E \to M$ such that observing $f(\tilde{e}_1, \tilde{e}_2)$ fully reveals $\tilde{e}_1$ and $\tilde{e}_2$. However, for any epsilon $\varepsilon \in (0, 1)$, there is a smooth map $f\colon E \to M$ such that the mean-squared error $E\left[(\tilde{e}_j - E\left[\tilde{e}_j \mid f(\tilde{e}_1, \tilde{e}_2)\right])^2\right]$ is less than ε for $j = 1, 2$.

Given the lack of measures of communication complexity, team theory has focused on the characterization of the individual decision rules (the strategy profile) that maximize the expected payoff, given a fixed game form. This is multi-person statistical decision theory. The simplest interesting class of problems is when there are n players who observe private signals about the state of nature and then simultaneously choose actions. (This is a Bayesian game in which players have the same payoffs for every realization of the state and every action profile.) The difference between this and a single-person static Bayesian decision problem is that in the latter the problem is to choose a plan (mapping from states to actions) subject to a single measurability constraint (the plan must be measurable with respect to the decision maker's information) and in the former each of the n plans must satisfy its own measurability constraint.

In this example, there is no communication. A richer example is the resource allocation model studied in Radner (1972), Groves and Radner (1972), Arrow and Radner (1979), and Groves (1983). The basic decision problem is the resource allocation problem (1) stated in Section 4, and the organizational structures typically resemble the decomposition given there, with a central production manager allocating resources and each unit manager controlling the unit's local decision variable. However, whereas the decomposition and message-space literatures were interested in the ability of an iterative procedure to asymptotically compute the optimal solution for any data in a specified domain, these team theory models consider fixed, finite exchanges of information that do not resolve all uncertainty before decisions are made. A small sample of other models is Beckmann (1958), Groves and Radner (1972), Marschak (1972), Cremer (1980), Aoki

(1986), and Green and Laffont (1986).

One paper, by Geanakoplos and Milgrom (1991), is different because the agents are drawn from a pool of managers with limited ability to acquire information about the environment. These constraints are motivated by the time it takes a manager to read and understand information. The decision problem is to allocate resources to a set of shops. An organization or team is a hierarchy whose leaves are the shops and whose interior nodes are managers selected from the pool. The managers recursively disaggregate the resource allocations, after acquiring information about the cost functions of the shops from external sources. (There is no flow of information up the hierarchy.) The value of decentralization, i.e., of hierarchies with more managers, is roughly that it allows more information to be used to allocate resources.

Because this paper models constraints on the ability to process raw data, it is closely related to the literature on decentralized computation reviewed in Section 6. However, the paper does not explicitly model individual constraints on calculations nor does it model information processing delay, both of which are emphasized by that literature. Furthermore, a basic part of the analysis is the characterization of the optimal decision rules for a fixed organization and information structure, which is a classic team-theory exercise.

5.5. Communication and incentive constraints

Incentive compatibility means that the agents' preferences over outcomes are given exogenously and the strategy profile that is part of a communication procedure must be a game-theoretic equilibrium. Hurwicz (1972) introduced incentive compatibility into the design of decentralized communication procedures, but the incentives-based mechanism design literature that followed has for the most part ignored communication costs. Most of it makes use of direct revelation mechanisms in which, in one round of simultaneous communication, all agents communicate all their private information. However, there is some research that combines incentive and communication constraints.

The simplest exercise is to add communication costs to a standard mechanism design problem, with the restriction to one-stage simultaneous-move mechanisms. The agents and their environments are defined as in Section 5.2. A mechanism is described by a message space M_i for each agent i and an outcome function $g\colon M \to A$, where again $M = \prod_{i=1}^{n} M_i$.[9] Exogenously given is the utility function $u_i\colon A \times E_i \to \mathbb{R}$ for each agent i. Each agent i observes either her own environment or the entire profile of environments, and announces (at the same time as the other agents) a message $m_i \in M_i$. If agents observe the entire profile of environments, the solution concept is Nash equilibrium. If they observe only their own environment, there is some probability measure on E and the solution concept is Bayesian-Nash equilibrium. If $\mu(e)$ is the set of equilibria when the environment is e, then the goal function implemented by the mechanism is $g \circ \mu\colon E \to A$.

The standard implementation problem is to take as given a goal function $F\colon E \to A$ and ask whether a mechanism exists that implements F. This can be modified by

9. In abstract mechanism design, M_i may be the strategic-form representation of a dynamic extensive form. However, the size of the strategic-form strategy space is not a good measure of the communication requirements of extensive forms. For example, suppose agent 1 announces a number, which is observed by agent 2, and then agent 2 announces a number. The strategy space of agent 1 is $\mathbb{R}$ and is one-dimensional, but that of agent 2 is $\mathbb{R}^{\mathbb{R}}$, and hence is infinite dimensional.

asking which of the mechanisms that implements F has the smallest messages space M. This problem was set up in Hurwicz (1972), and has been studied, for example, in Williams (1986), Reichelstein and Reiter (1988), and Hong and Page (1994), using Nash implementation.

The standard mechanism design problem introduces preferences of the mechanism designer (principal) over goal functions and characterizes the principal's preferred mechanism. This exercise can be modified by adding a cost to mechanisms that is increasing in the size of the message space or simply imposing a bound on the size of the message space, and then asking which mechanism is preferred by the principal taking into account both the goal function that is implemented and the communication cost. An example of such an exercise is Green and Laffont (1987), which also uses Nash equilibrium.

There is a close formal relationship between static Nash implementation and Hurwicz's static reduction of iterative procedures. Given a mechanism $\langle M, g\rangle$, we can write an agent's best reply function as $f_i\colon M\times E_i \to M_i$. Given an environment e, a Nash equilibrium in the game induced by the mechanism is a stationary message when we treat f_i as agent's i's response function in an iterative communication procedure with message-space M and outcome function g. Thus, if the mechanism $\langle M, g\rangle$ Nash-implements a goal function F, then $\left\langle M, g, \{f_i\}_{i=1}^n \right\rangle$ is an iterative procedure that evaluates F. The difference between the design of an iterative procedure that evaluates F and a mechanism that implements F is that, in the former, the response functions can be chosen by the procedure designer whereas in the latter, they are determined endogenously by the agents' utility maximization.

As I have presented them above, these exercises of static Nash implementation are very different in spirit from the communication problems described in Sections 5.2–5.4. They implicitly presume that all agents know each other's information and the goal function is computed, in the sense that its value is communicated to the mechanism designer, by a single exchange of information. However, the interpretation intended by these authors is that the Nash equilibria are the asymptotic steady states of an iterative procedure in which initially information is asymmetric. For this, we should imagine that trade or outcomes actually take place every period. This interpretation is partially supported by the large literature on learning in repeated games with initially asymmetric but fixed information in which adaptive or rational learning by agents can induce convergence to the symmetric information Nash equilibrium. The problem of incorporating incentive constraints in explicit dynamic iterative communication procedures with fixed information is outlined and explored in Roberts (1987), which emphasizes examples from the planning literature.

Another approach is to model the short-run revelation of information, as in mechanism design with asymmetric information between the agents, using dynamic extensive game forms. With networked communication, we can take into account individual communication requirements. In particular, the principal has very high communication requirements in static mechanisms because she must listen to all messages. These requirements may be reduced by mechanisms in which some agents communicate with each other but not with the principal, such as when a principal contracts with one agent, who then contracts with a second agent. Such mechanism design (but so far with somewhat ad-hoc communication constraints and cost) has been studied in Mookherjee and Reichelstein (1995, 1996), McAfee and McMillan (1995), Laffont and Martimort (1996), Baliga and Sjöström (1996), and Melumad et al. (1997), as well as in the

incomplete contracts literature.

6. Computation constraints

The communication complexity literature is based on the idea that agents are exogenously endowed with private information that is needed to compute a decision rule. More recently, economists have used models of decentralized computation, in which the computation constraints of individual agents are also modeled. Agents with no prior private information may thus be hired for the sole purpose of sharing information processing tasks. Hence, this is a good paradigm for studying endogenous administrative staffs.

In Section 6.1, I briefly outline the potential specifications of a model of decentralized or parallel computation. Section 6.2 reviews models in which the problem is to compute an exogenously given function, and Section 6.3 considers models of real-time computation, in which the problem is to compute an endogenous decision rule in a temporal decision problem.

6.1. Parallel and distributed processing

Simple models of joint computation by humans, which I refer to as decentralized computation, naturally resemble simple models of parallel or distributed computation, as joint computation by machines is called in computer science. A model of parallel, distributed or decentralized computation specifies (i) how each agent calculates and (ii) how agents communicate and coordinate their computation. The specification involves breaking down these tasks into steps that take up the agents' time, and also describing the memory capacities of the agents and the information transmission capabilities of the network that connects them.

Note that such a model can provide a richer and more detailed representation of communication than in the communication complexity models of Section 5.2, because it can specify the steps that an agent must take to send and receive messages. In particular, individual constraints on reading and interpreting raw data and messages can be included. In such a model, the distinction between individual communication constraints and information transmission costs is explicit. The latter can also be included and can be measured in the same way as in Section 5.2.

Computer scientists distinguish between parallel and distributed computation. Models of distributed computation emphasize communication costs—including transmission costs—and coordination problems due to asynchronous operation of the agents, whereas models of parallel computation do not. Thus, parallel computation models are simpler but are most suited to situations in which the processing agents are nearby and tightly coordinated, such as are the components of a multiprocessor computer; in contrast, models of distributed computation are more realistic when agents are loosely coordinated and spatially separated, such as are workstations in a wide-area computer network.

The simplest model of parallel computing is the PRAM (Parallel Random Access Machine). This model suppress all communication costs, including individual communication constraints, and memory constraints. Since agents' memories can be kept identical by instantaneous exchanges of messages, it is possible to imagine that the

different agents simultaneously manipulate the same memory bank. It is also usually assumed that all agents are identical, which means that they are capable of executing the same elementary operations (e.g., calculating the sum of two items in the memory and storing the result in the memory) with the same delays. It is then possible to describe an algorithm by simply listing the operations that are performed in each moment of time. It does not matter which operations are assigned to which agent (as long as no agent is assigned two operations at the same time), because all agents are identical and have equal access to the global memory. This idealized model is used in theoretical computer science because of its simplicity, and provides at least lower bounds on parallel complexity.

Of course, human organizations are more realistically modeled by distributed computation. Furthermore, models of parallel computation with no communication constraints are not always useful for studying organizational structure, because the flow of information is not determinate. Nevertheless, many of the models of decentralized computation used so far by economists do not incorporate sophisticated measures of transmission costs or the problems associated with asynchronicity, in order to focus on individual computation and communication delays.

Most models of computation in computer science are based on discrete data representations and operations, whereas most economic models have continuous variables and analytic functions. As long as the set of elementary operations is simple, it is easy to interpret, e.g., a model of fixed-precision arithmetic as a model of arithmetic with real numbers, and the former will approximate the latter. However, a general model of computation with real numbers requires smoothness restrictions to prevent "computation smuggling". Such a model is developed by Mount and Reiter (1990) and Reiter (1996).

6.2. Batch processing

In batch processing, there is a given function that must be computed. All inputs are available at the same time, and delay is measured by the time between when the computation begins and when the final answer is computed. In *serial* batch processing (where serial means with a single agent), complexity is measured mainly by this delay,[10] which measures both how long one has to wait for an answer and how long the single agent is busy (work or CPU time). With *parallel* batch processing, delay and processing costs are no longer the same, and parallelization entails a trade-off between them. The benefit of parallelization is that some operations can be performed concurrently and this reduces delay. The potential cost of parallelization is that it increases processing costs, e.g., due to costs of communication between the processing agents.

Economic models of batch processing have studied associative computation more than any other problem. This is because it is simple and because it has a natural tree structure that can be interpreted as the hierarchical structure of organizations. It is also a very important and prevalent class of computation problems, and includes complex tasks such as project selection or aggregation of cost functions.

The efficient algorithms for associative computation with a PRAM are illustrated in Figure 1 for finding a maximum. When computation begins, the data are stored in the memory. Each cycle, the data or previous results are assigned in pairs to agents, and each agent computes the maximum of the two assigned numbers. The answer is obtained in

10. And perhaps also, e.g., by memory requirements.

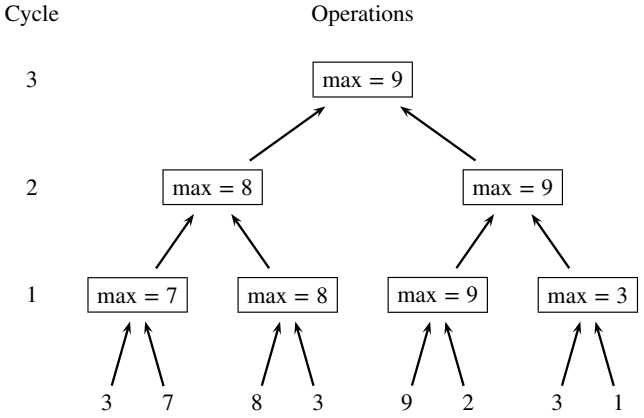

FIGURE 1. Associative computation by a PRAM via a balanced binary tree (Gibbons and Rytter (1988, Figure 1.2)).

$\lceil \log_2 n \rceil$ cycles, and there are $n-1$ operations. In contrast, if there were only a single agent, this agent would perform the $n-1$ operations consecutively and the delay would be $n-1$. In this case, with both serial and parallel processing, the processing cost, equal to the number of operations, is $n-1$. However, if there were communication costs, then the parallel algorithm would be more costly.

The computation is represented by a balanced binary tree in Figure 1, but note that the nodes are operations, rather than agents. Hence, we cannot interpret the tree in Figure 1 as the hierarchical structure of the organization. Furthermore, the flow of information is indeterminate with the PRAM, because there are no communication costs. In each cycle, the operations performed can be assigned arbitrarily to the agents. For example, in Figure 1, a agent might perform the left-most operation ($\max\{3,7\}$) in cycle 1, and then the right-hand operation ($\max\{9,3\}$) in cycle 2.

To make determinate the flow of information between individual agents, which the literature sometimes interprets as organizational structure, a model of computation with communication costs is needed. Radner (1993) uses a model of associative processing with individual communication constraints in which there is one additional operation for each message sent and processed. Radner (1993) characterizes the algorithms that are efficient with respect to delay and processing costs. Keren and Levhari (1979, 1983) can also be interpreted as the Radner model of associative computation, with the exogenous restriction that the communication networks be balanced hierarchies. However, although Radner (1993) finds that the efficient networks are hierarchical, they are not balanced and there is even skip-level reporting, which means that a manager can have subordinates with varying ranks. In fact, in the efficient hierarchies, all agents process the same amount of raw data.

Several new considerations arise when there is a flow of computation problems, as opposed to a one-shot computation problem. First, the workload has a new parameter—the throughput (problems per unit time). If agents are paid on an hourly basis and a different network can be used for each problem, then higher throughput can be handled just by setting up additional networks to process the problems. Under these assumptions, the computation of each problem is separate and should be performed the same way as in the one-shot case. However, if the agents are salaried, then they must be paid even when idle and the scheduling problem may lead to more regular hierarchies (Radner

(1993)). Nevertheless, Van Zandt (1997d) shows that the efficient networks are not even hierarchical and the processing of each problem closely resembles the efficient processing for the one-shot case. A stationarity restriction, meaning that each problem must be computed by the same network in the same way, has more significant consequences, because it is not possible to increase throughput simply by creating more networks. Bolton and Dewatripont (1994), using a generalized computation model, show that this restriction can lead to more regular hierarchies.

We can model the resource allocation problem as a batch processing problem. This provides an alternate approach to quantifying the communication and computational requirements of resource allocation processes, that is different from the measures that have been used in the decomposition, message-space and team theory literatures. Nemirovsky and Yudin (1983) and Ibaraki and Katoh (1988) provide a number of results on the complexity of resource allocation and other constrained optimization. However, Friedman and Oren (1995) were the first to study an algorithm that resembles decentralized communication procedures. They study the complexity of a price mechanism for classical economies. Rather than just studying stationary messages or convergence, they measure how many iterations are needed to achieve a given level of error. Furthermore, they model the calculations agents must perform to compute their messages. They calculate a bound on the run time as a function of the number n of agents and the error.

What was understood in the earlier literature on decomposition methods and communication procedures, but was not made explicit until Friedman and Oren (1995), is that the iterative procedures not only economize on communication, but they also decentralize computation tasks among the agents. In Friedman and Oren (1995), the center must still aggregate demands, the complexity of which increases with the number of agents, but at each iteration the agents are computing their own excess demands. If the agents instead sent their entire utility functions to the center, the center might still use the same iterative algorithm, but the computational load of the center would greatly increase.

Other models of decentralized batch processing include Hong and Page (1995) and Li (1997). Queueing models with stochastic arrivals of computation tasks include Malone and Smith (1988) and Beggs (1995). As all these authors consider problems other than associative computation, the hierarchical structure that is a natural part of associative computation is sometimes not present. This is realistic, as information flows in organizations are typically not fully hierarchical even if there is a hierarchical structure of authority. Jordan (1995) is a model of hierarchies, but they are decision trees. His paper studies the evolution of organizational structure. Miller (1996) also drops the constrained-optimal approach and instead studies the evolution by genetic algorithms of associative computation hierarchies.

6.3. Real-time computation

In the previous section, I noted that computational delay is an important cost (measure of complexity) of computation. Furthermore, the importance of delay in organizations has long been discussed in the economics literature. For example, Kaldor (1934, p. 78) observed that coordination tasks arise only in changing, dynamic environments, and Robinson (1958, Chapter III) emphasized managerial delay as a limit to firm size. Hayek (1940, pp. 131-132), in a criticism of the iterative adjustment processes proposed by

Lange (1936, 1937) and Dickinson (1939), which assume that the underlying economic data are constant, states:

> In the real world, where constant change is the rule, … whether and how far anything approaching the desirable equilibrium is ever reached depends entirely on the speed with which the adjustments can be made. The practical problem is not whether a particular method would eventually lead to a hypothetical equilibrium, but which method will secure the more rapid and complete adjustment to the daily changing conditions ….

However, delay and its affect on organizations are not well captured by the models of batch processing described in the previous subsection.

Delay is costly in decision problems because it increases the lag upon which decisions are based. To capture this cost, we need to model the real-time computation of dynamic decision problems, i.e., the "adjustment to the daily changing conditions". This means that we begin with a temporal decision problem in which information becomes available over time and a sequence of decisions is made. The computation of decision rules must then adapt itself to the dynamic structure of the decision problem.

Marschak (1972) studies several real-time adjustment processes for allocating resources and explores some of the costs of delay, although the computation is not explicitly modeled. Real-time parallel computation was first used in the theory of organizations by Radner and Van Zandt (1992) and Van Zandt and Radner (1997), for a problem of predicting the sum of a family of stochastic processes. The purpose of these papers is to characterize returns to scale.

The results in Van Zandt and Radner (1997) illustrate how replication arguments for deriving non-decreasing returns to scale of firms break down when information processing is modeled. By replication arguments, I mean the claim that returns to scale should be non-decreasing because a large firm can mimic the activities of several small firms. Intuitively, such arguments are not valid because mimicking the activities of several small firms requires that the subunits not communicate, coordinate their activities, or allocate resources except as independent firms would do, such as through markets. There can be no headquarters that controls the subunits because such control would incur managerial costs that the independent firms avoid. It is hard to envision that such informationally disintegrated units constitute a single firm.[11]

Van Zandt and Radner (1997) use a replication argument to show that returns to scale are eventually increasing, under certain statistical assumptions, when the sampling of data is costly but information processing is instantaneous. A firm that must predict the sum of kn processes can divide the processes into k groups or divisions of size n, and each division can imitate the sampling and prediction policies of a firm with n processes. The overall prediction is the sum of the divisions' k predictions. This argument does not work when computation of decision rules is modeled. If the firm divides itself into k divisions that imitate the computation procedures of a firm of size n, then the result is k predictions that must be aggregated. Because of the aggregation delay, the prediction of the firm of size kn uses data that is older than the prediction of a firm of size n.

11. However, a theory of how informational integration is related to the boundaries of the firm, combining decentralized information processing and the property-rights theory of the firm (see, e.g., Hart (1995)), has not yet been developed.

Van Zandt (1997a) uses the same decision problem as in Radner and Van Zandt (1992) and Van Zandt and Radner (1997) to explore some of the special properties of real-time computation. For example, with batch processing, there is an unequivocal reduction in delay that results from increased parallelization. However, with real-time processing, each decision may use data of heterogeneous lags and hence there is no single measure of delay. The speed-up from parallelization is ambiguous, because when processing a report sent by another agent, an agent forgoes processing raw data that has a lower lag than all the data that went into the report. Van Zandt (1997d) present an example where there is no decentralization because of this decision-theoretic cost, even if the managerial wage is zero and hence decentralization has no resource cost.

Real-time computation is used in Van Zandt (1997b, 1997c, 1997e) for the problem of allocating resources. These papers describes a class of hierarchical allocation procedures in which information about the operatives' cost parameters flows up and is aggregated by the hierarchy, while allocations flow down and are disaggregated by the hierarchy. Multilevel hierarchies and decentralized decision making arise due to computational delay, rather than communication constraints. Offices at the bottom of the hierarchy allocate resources to a small number of shops and can thus compute their allocations quickly, while offices at higher levels use older information but can still make advantageous transfers to the subdivisions. Hence, the model illustrates how decentralization of hierarchical decision making can arise purely because of computational delay, even when there are no communication costs, all agents are identical and have the same free access to data about the environment, and there are no incentive problems.

7. Down the road

The bounded rationality and the incentive approaches to organizations have so far developed mainly independently, because of the advantages of first studying isolated phenomena. However, it has always been recognized that incentive problems in organizations are tightly bound to information processing. Contracting problems in management only arise because bounds on managers' information processing capacity lead to delegation of managerial tasks. The incompleteness of contracts that has been important in the theory of organizations is loosely motivated by bounded rationality. Furthermore, as studied in Aghion and Tirole (1995) and Chang and Harrington (1996), organizational structure can affect the incentives of agents to acquire and process information, because the structure affects the influence agents have over the organization's decisions. The integration of communication complexity and mechanism design was reviewed in Section 5.5. That area of research is still in a preliminary stage. Furthermore, there has been no integration of formal models of decentralized computation with contracting and mechanism design with incentive constraints.

Bounded rationality within organizations is also relevant to inter-organizational strategic interaction. The models of organizations surveyed here can be the basis for modeling boundedly rational players (where the players are firms or other types of organizations) in games. The real-time learning models are particularly suitable. For example, Meagher (1996) is a real-time decentralized information processing model in which the decision problem is to choose location or product characteristics when launching new products, given changing preferences of consumers. One can imagine

embedding this into an oligopoly model with differentiated products.

A question that has eluded satisfactory general treatment is the comparison of the computational efficiency of the many organizational forms, including market mechanisms and bureaucratic structures, that are used to coordinate economic activity. This question was posed by the markets versus planning and markets versus firms debates discussed in Section 3. This survey has concentrated on information processing models of various bureaucratic structures. There is also a large and rapidly growing literature on the dynamics of decentralized interaction between agents who are modeled as computing machines, including the literature on agent-based computational economics (see Tesfatsion (1997) for a survey) and multi-agent systems (e.g., Youssefmir and Huberman (1995)). However, the agents are not modeled in a way that would easily allow the representation of and comparison with bureaucratic procedures.

Appendix: Related research in other fields

As noted in the introduction, I have not attempted to give a proper interdisciplinary review of the theory of decentralized information processing in human organizations. As a slight compensation, I will briefly mention some of the fields that were skipped. Much of this research is more applied than what was reviewed in the body of the paper, although I have still omitted the empirical research from all disciplines. Note that the boundaries between the areas described below are not sharp.

In the fields of distributed computation and distributed artificial intelligence, there is a new awareness that the interaction between machines and also the human participants in distributed computer systems involves problems that economists have long studied, such as the allocation of scarce resources, the dynamics of semi-coordinated interactions between agents, and incentives of autonomous agents. For example, Waldspurger et al. (1992), Cheng and Wellman (1996), and Lee and Durfee (1995) study price-based resource allocation mechanisms related to those reviewed in this survey and Sandholm and Lesser (1995) include incentive constraints in distributed systems. Huberman (1996) reviews other relationships between economics and distributed systems.

Organization theory from a management and sociological perspective has a long history of studying human organizations as information processing networks (e.g., March and Simon (1958), Simon (1976), and Galbraith (1977)). Recent research incorporates explicit models of computation and increased use of mathematical modeling and simulation methods for studying the performance and evolution of organizations. A sample can be found in Carley and Prietula (1994); see Carley (1995) for a review.

The field of management and information systems also contains substantial research on information processing by humans and machines in organizations. Some of this falls under the new name of "coordination theory". See, for example, Malone (1987). For a more interdisciplinary review of this and related areas of applied organization theory, see Van Alstyne (1997).

References

Abselon, H. (1980). Lower bounds on information transfer in distributed computations. *Journal of the ACM*, 27, 384–392. First published in 1978.

Aghion, P. and Tirole, J. (1995). Formal and real authority in organizations. Nuffield College, Oxford and IDEI, Toulouse.

Aoki, M. (1986). Horizontal vs. vertical information structure of the firm. *American Economic Review*, 76, 971–983.

Arrow, K. and Hurwicz, L. (1960). Decentralization and computation in resource allocation. In R. W. Pfouts (Ed.), *Essays in Economics and Econometrics* (pp. 34–104). Chapel Hill: University of North Carolina Press.

Arrow, K. J. and Hurwicz, L. (Eds.). (1977). *Studies in Resource Allocation*. Cambridge: Cambridge University Press.

Arrow, K. J., Hurwicz, L., and Uzawa, H. (Eds.). (1958). *Studies in Linear and Non-Linear Programming*. Stanford: Stanford University Press.

Arrow, K. J. and Radner, R. (1979). Allocation of resources in large teams. *Econometrica*, 47, 361–385.

Baliga, S. and Sjöström, T. (1996). Decentralization and collusion. Department of Applied Economics, Cambridge University and Department of Economics, Harvard University.

Barone, E. (1935). The ministry of production in the collectivist state. In F. A. v. Hayek (Ed.), *Collectivist Economic Planning* (pp. 245–290). London: George Routledge and Sons. Originally published in 1908.

Baumol, W. J. and Fabian, T. (1964). Decomposition, pricing for decentralization and external economies. *Management Science*, 11, 1–32.

Beckmann, M. (1958). Decision and team problems in airline reservations. *Econometrica*, 26, 134–145.

Beggs, A. W. (1995). Queues and hierarchies. Wadham College, Oxford University.

Binmore, K. and Samuelson, L. (1992). Evolutionary stability in repeated games played by finite automata. *Journal of Economic Theory*, 57, 278–305.

Bolton, P. and Dewatripont, M. (1994). The firm as a communication network. *Quarterly Journal of Economics*, 109, 809–839.

Burton, R. M. and Obel, B. (1984). *Designing Efficient Organizations: Modelling and Experimentation*. Amsterdam: North-Holland.

Carley, K. (1995). Computational and mathematical organization theory: Perspective and directions. *Computational and Mathematical Organization Theory*, 1, 39–56.

Carley, K. M. and Prietula, M. J. (Eds.). (1994). *Computational Organization Theory*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Chandler, A. D. (1966). *Strategy and Structure*. New York: Doubleday.

Chandler, A. D. (1990). *Scale and Scope: The Dynamics of Industrial Capitalism*. Cambridge, MA: Harvard University Press.

Chang, M.-H. and Harrington, J. E. (1996). Organizational structure and firm innovation. Cleveland State University and John Hopkins University.

Cheng, J. Q. and Wellman, M. P. (1996). The WALRAS algorithm: A convergent distributed implementation of general equilibrium outcomes. AI Laboratory, University of Michigan.

Coase, R. (1937). The nature of the firm. *Economica*, 4, 386–405.

Cremer, J. (1980). A partial theory of the optimal organization. *The Bell Journal of Economics*, 11, 683–693.

Dantzig, G. B. and Wolfe, P. (1960). Decomposition principles for linear program. *Operations Research*, 8, 101–111.

Dickinson, H. D. (1939). *Economics of Socialism*. Oxford: Oxford University Press.

Dirickx, Y. M. I. and Jennergren, L. P. (1979). *Systems Analysis by Multilevel Methods*. Chichester, England: John Wiley and Sons.

Friedman, E. J. and Oren, S. S. (1995). The complexity of resource allocation and price mechanisms under bounded rationality. *Economic Theory*, 6, 225–250.

Fudenberg, D. and Tirole, J. (1991). *Game Theory*. Cambridge, MA: MIT Press.

Galbraith, J. (1977). *Organization Design*. Reading, MA: Addison-Wesley.

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

Gibbons, A. and Rytter, W. (1988). *Efficient Parallel Algorithms*. Cambridge: Cambridge University Press.

Green, J. and Laffont, J.-J. (1986). Alternative limited communication systems. In W. Heller, R. Starr, and D. Starett (Eds.), *Uncertainty, Information and Communication (Essays in Honor of K. J. Arrow, v. 3)*. Cambridge: Cambridge University Press.

Green, J. and Laffont, J.-J. (1987). Limited communication and incentive constraints. In T. Groves, R. Radner, and S. Reiter (Eds.), *Information, Incentives, and Economic Mechanisms*. Minneapolis: University of Minnesota Press.

Groves, T. (1983). The usefulness of demand forecasts for team resource allocation in a dynamic environment. *Review of Economic Studies*, 50, 555–571.

Groves, T. and Radner, R. (1972). Allocation of resources in teams. *Journal of Economic Theory*, 4, 415–441.

Hart, O. (1995). *Firms, Contracts, and Financial Structure*. Oxford: Oxford University Press.

Hayek, F. A. v. (1935). The nature and history of the problem. In F. A. v. Hayek (Ed.), *Collectivist Economic Planning* chapter 1. London: George Routledge and Sons.

Hayek, F. A. v. (1940). Socialist calculation: The competitive 'solution'. *Economica*, 7, 125–149.

Heal, G. (1986). Planning. In K. J. Arrow and M. D. Intriligator (Eds.), *Handbook of Mathematical Economics*, volume III chapter 29. Amsterdam: North Holland.

Heal, G. M. (1973). *The Theory of Economic Planning*. Amsterdam: North-Holland Publishing Co.

Hong, L. and Page, S. (1994). Reducing informational costs in endowment mechanisms. *Economic Design*, 1, 103–117.

Hong, L. and Page, S. (1995). Computation by teams of heterogeneous agents. Syracuse University and California Institute of Technology.

Huberman, B. A. (1996). Computation as economics. Dynamics of Computation Group, Xerox Palo Alto Research Center.

Hurwicz, L. (1960). Optimality and informational efficiency in resource allocation processes. In K. Arrow, S. Karlin, and P. Suppes (Eds.), *Mathematical Methods in the Social Sciences*. Stanford: Stanford University Press.

Hurwicz, L. (1972). On informationally decentralized systems. In C. B. McGuire and R. Radner (Eds.), *Decision and Organization*. Amsterdam: North-Holland Publishing Co. Second edition published in 1986 by University of Minnesota Press.

Hurwicz, L. (1977). On the dimensionality requirements of informationally decentralized pareto-satisfactory processes. In K. J. Arrow and L. Hurwicz (Eds.), *Studies in Resource Allocation Processes*. Cambridge: Cambridge University Press.

Hurwicz, L. (1986). On informational decentralization and efficiency in resource allocation mechanisms. In S. Reiter (Ed.), *Studies in Mathematical Economics*. Providence: The Mathematical Association of American. MAA Studies in Mathematics, vol. 25.

Ibaraki, T. and Katoh, N. (1988). *Resource Allocation Problems: Algorithmic Approaches*. Boston: MIT Press.

Jordan, J. (1987). The informational requirements of local stability in decentralized allocation mechanisms. In T. Groves, R. Radner, and S. Reiter (Eds.), *Information, Incentives, and Economic Mechanisms*. Minneapolis: University of Minnesota Press.

Jordan, J. (1995). Classification dynamics in the theory of decisions and organizations. University of Minneapolis.

Kaldor, N. (1934). The equilibrium of the firm. *Economic Journal*, 44, 70–71.

Keren, M. and Levhari, D. (1979). The optimum span of control in a pure hierarchy. *Management Science*, 11, 1162–1172.

Keren, M. and Levhari, D. (1983). The internal organization of the firm and the shape of average costs. *The Bell Journal of Economics*, 14, 474–486.

Kushilevitz, E. and Nisan, N. (1997). *Communication Complexity*. Cambridge: Cambridge University Press.

Laffont, J.-J. and Martimort, D. (1996). Collusion and delegation. IDEI, Université de Toulouse I.

Lange, O. (1936). On the economic theory of socialism: Part one. *Review of Economic Studies*, 4, 53–71.

Lange, O. (1937). On the economic theory of socialism: Part two. *Review of Economic Studies*, 4, 123–142.

Lee, J. and Durfee, E. H. (1995). A microeconomic approach to intelligent resource sharing in multiagent systems. Technical Report CSE-TR-234-95, AI Laboratory, University of Michigan.

Li, H. (1997). Hierarchies and information processing organizations. University of Chicago.

Malone, T. W. (1987). Modeling coordination in organizations and markets. *Management Science*, (33), 1317–1332.

Malone, T. W. and Smith, S. A. (1988). Modeling the performance of organizational structures. *Operations Research*, 36, 421–436.

March, J. G. and Simon, H. A. (1958). *Organizations*. New York: Wiley.

Marschak, J. (1955). Elements for a theory of teams. *Management Science*, 1, 127–137.

Marschak, J. and Radner, R. (1972). *Economic Theory of Teams*. New Haven: Yale Univeristy Press.

Marschak, T. (1972). Computation in organizations: The comparison of price mechanisms and other adjustment processes. In C. B. McGuire and R. Radner (Eds.), *Decision and Organization* chapter 10, (pp. 237–281). Amsterdam: North-Holland Publishing Co. Second edition published in 1986 by University of Minnesota Press.

Marschak, T. (1986). Organizational design. In K. J. Arrow and M. D. Intriligator (Eds.), *Handbook of Mathematical Economics*, volume III chapter 27, (pp. 1358–1440). Amsterdam: Elsevier Science Publishers.

Marschak, T. and Reichelstein, S. (1995). Communication requirements for individual agents in networks and hierarchies. In J. Ledyard (Ed.), *The Economics of Informational Decentralization: Complexity, Efficiency and Stability*. Boston: Kluwer Academic Publishers.

Marschak, T. and Reichelstein, S. (1996). Network mechanisms, informational efficiency, and hierarchies. Haas School of Business, University of California, Berkeley.

McAfee, R. P. and McMillan, J. (1995). Organizational diseconomies of scale. University of Texas (Austin) and University of California (San Diego).

Meagher, K. J. (1996). How to chase the market: An organizational and computational problem in decision making. Australian National University.

Melumad, N., Mookherjee, D., and Reichelstein, S. (1997). Contract complexity, incentives and the value of delegation. *Journal of Economics and Management Strategy*, 6.

Mesarovic, M. D. and Takahara, Y. (1989). *Abstract Systems Theory*. Berlin: Springer-Verlag.

Miller, J. (1996). Evolving information processing organizations. Carnegie Mellon University.

Mises, L. v. (1951). *Socialism: An Economic and Sociological Analysis*. New Haven: Yale University Press. Originally published as *Die Gemeinwirtschaft* in 1922.

Mookherjee, D. and Reichelstein, S. (1995). Incentives and coordination in hierarchies. Boston University and Haas School of Business (UC Berkeley).

Mookherjee, D. and Reichelstein, S. (1996). Budgeting and hierarchical control. Boston University and Haas School of Business (UC Berkeley).

Moore, J. C., Rao, H. R., and Whinston, A. B. (1996). Information processing for finite resource allocation mechanisms. *Economic Theory*, 8, 267–290.

Mount, K. and Reiter, S. (1974). The informational size of the message space. *Journal of Economic Theory*, 8, 161–192.

Mount, K. and Reiter, S. (1977). Economic environments for which there are pareto satisfactory mechanisms. *Econometrica*, 45, 821–842.

Mount, K. and Reiter, S. (1987). The existence of a locally stable dynamic process with a statically minimal message space. In T. Groves, R. Radner, and S. Reiter (Eds.), *Information, Incentives, and Economic Mechanisms*. Minneapolis: University of Minnesota Press.

Mount, K. and Reiter, S. (1990). A model of computing with human agents. The Center for Mathematical Studies in Economics and Management Science, Discussion Paper No. 890, Northwestern University, Evanston, Illinois.

Nemirovsky, A. S. and Yudin, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. New York: Wiley.

Osborne, M. and Rubinstein, A. (1994). *A Course in Game Theory*. Cambridge, MA: MIT Press.

Pareto, V. (1927). *Manuel d'Economie Politique*. Paris: Marcel Giard.

Radner, R. (1962). Team decision problems. *Annals of Mathematical Statistics*, 33, 857–881.

Radner, R. (1972). Allocation of a scare resource under uncertainty: An example of a team. In

C. B. McGuire and R. Radner (Eds.), *Decision and Organization* chapter 11, (pp. 217–236). Amsterdam: North-Holland Publishing Co. Second edition published in 1986 by University of Minnesota Press.

Radner, R. (1993). The organization of decentralized information processing. *Econometrica*, 62, 1109–1146.

Radner, R. and Van Zandt, T. (1992). Information processing in firms and returns to scale. *Annales d'Economie et de Statistique*, 25/26, 265–298.

Reichelstein, S. and Reiter, S. (1988). Game forms with minimal message spaces. *Econometrica*, 56, 661–700.

Reiter, S. (1977). Information and performance in (new)^2 welfare economics. *American Economic Review: Paper and Proceedings*, 77, 226–234.

Reiter, S. (1996). Coordination and the structure of firms. Northwestern University.

Reiter, S. and Simon, C. (1992). A decentralized dynamic process for finding equilibrium. *Journal of Economic Theory*, 56, 400–425.

Roberts, J. (1987). Information, incentives and iterative planning. In T. Groves, R. Radner, and S. Reiter (Eds.), *Information, Incentives, and Economic Mechanisms*. Minneapolis: University of Minnesota Press.

Robinson, A. (1934). The problem of management and the size of firms. *Economic Journal*, 44, 240–254.

Robinson, E. A. G. (1958). *The Structure of Competitive Industry*. Chicago: The University of Chicago Press.

Rogers, D. F., Plante, R. D., Wong, R. T., and Evans, J. R. (1991). Aggregation and disaggregation techniques and methodology in optimization. *Operations Research*, 39, 553–582.

Sandholm, T. and Lesser, V. R. (1995). Equilibrium analysis of the possibilities of unenforced exchanged in multiagent systems. In *Proceedings 14th International Joint Conference on Artificial Intelligence*. Montreal.

Segal, I. (1996). Communication complexity and coordination by authority. Department of Economics, University of California at Berkeley.

Sethi, S. P. and Zhang, Q. (1994). *Hierarchical Decision Making in Stochastic Manufacturing Systems*. Boston: Birkhaüser Boston.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379–423, 623–656.

Simon, H. A. (1976). *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations*. New York: Free Press.

Taylor, F. M. (1929). The guidance of production in the socialist state. *American Economic Review*, 19.

Tesfatsion, L. (1997). How economists can get alife. In W. B. Arthur, S. Durlauf, and D. Lane (Eds.), *The Economy as an Evolving Complex System, II*. Menlo-Park: Addison-Wesley.

Van Alstyne, M. (1997). The state of network organization: A survey in three frameworks. *Journal of Organizational Computing*. Forthcoming.

Van de Panne, C. (1991). Decentralization for multidivision enterprises. *Operations Research*, 39, 786–797.

Van Zandt, T. (1997a). Real-time decentralized information processing as a model of organizations with boundedly rational agents. *Review of Economic Studies*. Forthcoming.

Van Zandt, T. (1997b). Real-time hierarchical resource allocation. Princeton University.

Van Zandt, T. (1997c). Real-time hierarchical resource allocation with quadratic costs. Princeton University.

Van Zandt, T. (1997d). The scheduling and organization of periodic associative computation: I Essential networks, II Efficient networks. *Economic Design*. Forthcoming.

Van Zandt, T. (1997e). Structure and returns to scale of real-time hierarchical resource allocation. Princeton University.

Van Zandt, T. and Radner, R. (1997). Real-time decentralized information processing and returns to scale. Princeton University and New York University.

Waldspurger, C. A., Hogg, T., Huberman, B. A., Kephart, J. O., and Stornetta, W. S. (1992). Spawn: A distributed computational economy. *IEEE Transactions on Software Engineering*, 18, 103–117.

Walras, L. (1954). *Elements of Pure Economics*. London: George Allen and Unwin. Translated by W. Jaffé.

Williams, S. R. (1986). Realization and Nash implementation: Two aspects of mechanism design. *Econometrica*, 54, 139–151.

Williamson, O. E. (1975). *Markets and Hierarchies, Analysis and Antitrust Implicatons*. New York: Free Press.

Williamson, O. E. (1985). *The Economic Institutions of Capitalism*. New York: Free Press.

Yao, A. C. (1979). Some complexity questions related to distributive computing. In *Proceedings of the 11th ACM Symposium on Theory of Computing* (pp. 209–213).

Youssefmir, M. and Huberman, B. A. (1995). Clustered volatility in multiagent dynamics. Dyamics of Computation Group, Xerox Palo Alto Research Center.