# Economics, Organization and Management

Paul Milgrom
*Stanford University*

John Roberts
*Stanford University*

Prentice Hall, Englewood Cliffs, New Jersey 07632

# 2

# ECONOMIC ORGANIZATION AND EFFICIENCY

*[People] in general, and within limits, wish to behave economically, to make their activities and their organization "efficient" rather than wasteful.*

*Frank Knight*[1]

This book is concerned with the problems of designing and managing efficient economic organizations. The preceding chapter gives examples of the range of problems and institutions that we are considering. We now explore the source and nature of these problems and the approach we take to analyzing them.

## ECONOMIC ORGANIZATIONS: A PERSPECTIVE

Economic organizations are created entities within and through which people interact to reach individual and collective economic goals. The economic system consists of a network of people and organizations with lower-level organizations linked together through higher-level organizations.

Types of organizations:
- lower-level } organizations
- higher-level }

The highest-level organization is the economy as a whole. While it is somewhat unusual to think of an entire economy as an organization, this perspective is useful because it emphasizes that the economic system is a human creation and because many of the problems that smaller, more formal organizations face exist at the economy-wide level as well. As an organization, an economy can and should be

[1] "Review of Melville J. Herskovits' 'Economic Anthropology,'" *Journal of Political Economy*, 49 (April 1941), 246–258, quoted by Oliver Williamson in "Mergers, Acquisitions, and Leveraged Buyouts: An Efficiency Assessment," Working Paper Series D, Economics of Organization, no. 30, Yale University School of Organization and Management (1987). (Emphasis added by Williamson.)

evaluated based on its performance relative to possible alternative arrangements. The current experiments in thoroughgoing reform of entire economic systems in Eastern Europe are prime examples of the importance of this perspective.

have selfish,
ods, can do
etc...

## Formal Organizations

At the next level are the entities more traditionally regarded as organizations and the ones that are our main concern: corporations, partnerships, sole proprietorships, labor unions, government agencies, universities, churches, and other formal organizations. A key characteristic of the organizations at this level is their independent legal identity, which enables them to enter binding contracts, to seek court enforcement of those contracts, and to d[illegible] in their own name, separate from the individuals who belong to the organization.

Capacidade de formar contratos bilaterais com seus stakeholders

Organizations and Contracting This ability to enter contracts is critical to one of the major approaches to the economic analysis of organizations. In this view, which was first suggested by Armen Alchian and Harold Demsetz, an organization is regarded as a nexus of contracts, treaties, and understandings among the individual members of the organization. The firm itself is then a legal fiction that enters relatively simple, bilateral contracts between itself and its suppliers, workers, investors, managers, and customers. Without a legal entity that can contract with them individually, these people would have to fashion complex, multilateral agreements among themselves to achieve their aims.

The contracting approach to organization theory emphasizes the voluntary nature of people's involvement in (most) organizations: People will give their allegiance only to an organization that serves their interests. Furthermore, along with the ability to enter contracts come the possibilities for reform, redesign, and abandonment of the organization by rearranging contractual terms. This approach also facilitates accepting the fuzziness of organizational boundaries and the fact that organizational forms blend together. Markets and hierarchies—sometimes regarded as the major discrete alternative ways of organizing economic activity—are actually just two extreme forms of organizational contracting, with voluntary bargaining characterizing markets and strict lines of authority characterizing hierarchy.

The Architecture of Organizations Although the legal aspects of organization are important, a full description of organizational architecture involves many more elements: the patterns of resource and information flows, the authority and control relationships and the distribution of effective power, the allocation of responsibilities and decision rights, organizational routines and decision-making processes, the methods for attracting and retaining members and resources, the means by which new ideas and knowledge are generated and diffused throughout the organization, the adaptation of the organization's routines to reflect and implement organizational learning, the organization's expressed objectives and the strategies and tactics employed, and the means used to unifv the goals and behavior of the individual members of the organization and the objectives of the organization as a whole. Various parts of our analysis focus on each of these and on how the pieces fit together to yield a coherent pattern.

Once our focus becomes the elements of organizational architecture, defining a formal organization simply by its ability to contract as a distinct legal entity can become quite inappropriate because it can easily misidentify the effective boundaries of the organization. Consider, for example, the Sony company, which is known for its innovative consumer electronics products, but also manufactures broadcast equipment, computer components, and semiconductors and owns one of the three largest record companies in the world (Sony Records, formerly CBS Records) and a major movie and television production and distribution operation (Columbia Pictures Entertainment). Sony in fact consists of the Sony Corporation, the parent organization based in Tokyo, plus subsidiary corporations around the world. Each of these subsidiaries is a separate legal entity able to enter contracts on its own, and being an employee of Sony Corporation of America or of Sony GmbH in Germany does not make one an employee of Sony Corporation. Thus, the legal-entity approach might point to viewing each subsidiary as a separate organization. Yet this accords neither with the way the world sees Sony nor with the way it sees itself and manages its affairs.

boundaries: limites

Discretion and Autonomy from Intervention In these circumstances, a useful way to look at the defining boundaries of an organization is in terms of the smallest unit that is functionally autonomous in that it is largely free from intervention by outside parties in its affairs and decisions, over which it then enjoys broad internal discretion. Within a firm, the rightful decision makers—usually senior management—collectively have broad legal rights to order that activities be conducted as they see fit and to require that their directives be followed. Private outside parties cannot countermand these orders. Courts and government regulators may be able to intervene in some ways, but the discretion of even these public agencies is generally limited. They cannot interfere except under prescribed conditions, and then the measures they can take are limited compared to those that are available within the firm.

unidade geradora de caixa= UGC

Using this approach, the companies that make up Sony constitute a single organization, even though they are separate legal entities. The senior managers of Sony Corporation have the power to intervene largely as they wish in the operations of the subsidiaries, even if they in fact rarely choose to do so. Thus, the subsidiaries are not separate organizations by this test. Meanwhile, no outside private party has the legal right to specify how Sony will be run, what products it will make, what prices it will charge, what pay and job assignments it will offer, how it will divide its activities among its subunits, what investments it will make, and so on. Thus, Sony stands as a separate organization.

## The Level of Analysis: Transactions and Individuals

The most fundamental unit of analysis in economic organization theory is the **transaction**—the transfer of goods or services from one individual to another. The way a transaction is organized depends on certain of its characteristics. For example, if one kind of transaction occurs frequently in similar ways, people develop routines to manage it effectively. If a transaction is unusual, then the parties may need to bargain about its terms, which raises the costs of carrying out the transaction.

The ultimate participants in transactions are individual human beings, and their interests and behavior are of fundamental importance for understanding organizations. People are fundamental first in the sense of being indivisible decision makers and actors; it is people—not organizations—who actually decide, vote, or act. The actions of individuals determine the behavior and performance of organizations. Furthermore, only the needs, wants, and objectives of individuals have ethical significance. Economic organizations are judged only on the basis of how well they serve people's intended purposes. Finally, it is people who ultimately create and manage organizations, judge their performance, and redesign or reject them if this performance is found inadequate.

In analyzing how organizations emerge, how they are structured, how they function, and how economic activity is divided among them, we adopt the position put forward in the opening quotation: that people will seek to achieve efficiency in more than just the day-to-day conduct of their economic affairs. Efficiency also must exist at a systemic level, in the organization of people's activities and in the design, ficiency

y: no available choice/option
one in the relevant
ou prefers.

cy:
there is an alterna
sible choice
~~ld~~ would help
rson without
g any other"

management, and governance of the institutions they create. Before we explore this position more fully, we need to be more precise about what we mean by *efficiency*.

## EFFICIENCY

The goal of any economic organization, including the economic system as a whole, is to satisfy the wants and needs of individual human beings. We judge economic performance in terms of this goal. This approach does not imply an exclusive concern with materialistic achievement. If military dominance or national prestige is the priority, then an economy that serves these goals could be performing well by our standards. If the population were united in believing that the purpose of all human activity should be to glorify the deity, then an economic system that supported this consecration would be a good one. For purposes of our discussion, however, we assume that people primarily are concerned with regular economic goods and services. The economic system then is judged on how well it satisfies the economic needs of the population.

This approach obviously requires that we ascribe preferences to individuals. Indeed, we assume that people are equipped with measures of their welfare (called *utility functions*), that they like one situation better than another if and only if it gives greater utility, and that their economic goal is to maximize this measure of satisfaction. We explore this assumption in more detail later in the chapter.

The problem of scarcity, however, means that trade-offs typically have to be made. Increasing the utility of one person may mean having to give less to another. How then are we to measure performance? What does it mean to consider how well peoples' interests are served when these interests are possibly in conflict?

### The Concept of Efficiency

The partial solution is to focus on **efficient** choices or options, by which we mean ones for which *there is no available alternative that is universally preferred in terms of the goals and preferences of the people involved.* More precisely, if individuals are sometimes indifferent about some of the available options, then a choice is efficient if there is no other available option that everyone in the relevant group likes as least as much and at least one person strictly prefers. Turning the definition around, a choice is *inefficient* when there is an alternative possible choice that would help one person without harming any other.

Note well that the efficiency criterion can never be applied to resolve ethical questions about when it is justified or worthwhile to help one person at another's expense. In some situations, such questions cannot be avoided, but efficiency will not help. Instead, appeals to other criteria that explicitly trade off one individual's welfare against another's are needed.

Note too that the efficiency or inefficiency of a choice is always relative to some specific set of individuals whose interests are being taken into account and also to some specific set of available options. This is important to remember. It is distinctly possible that a particular choice from a given set of alternatives will be efficient relative to the interests of a given group of people, but not when some larger affected group is considered. Similarly, a choice may be efficient when all the constraints that delimit the set of available options are recognized, but not when the removal of some of these makes more options available. Thus, in applying the concept of efficiency it is necessary to be clear about whose interests are counted and what alternatives are considered to be feasible.

a eficiência está intimamente ligada às preferências individuais ou seja eficiência depende de outras variáveis

Pareto Dominated

Pareto Optimal

Pareto efficiency

eficiência de Pareto

### Efficiency of Resource Allocations

Efficiency can be defined and applied at many levels, depending on the kind of choice being considered. Our first application—and the most common one in economic analysis—is to compare alternative allocations of resources. An allocation of resources A is inefficient if there is some other available allocation B that everyone concerned likes at least as well as A and that one person strictly prefers. (In this case the allocation A is sometimes said to be *Pareto dominated*[2] by allocation B.) An inefficient allocation is wasteful: by making better use of the available resources it would be possible to make some people better off without hurting anyone else. If, on the other hand, no allocation exists that is unanimously preferred to A, then the given allocation A is efficient (or *Pareto optimal*). An efficient allocation of resources is thus one such that there is no other available allocation that makes someone better off without making another person worse off.

Although the reasons for wanting an efficient allocation of resources are obvious, efficiency by itself is a weak performance criterion. First, there are typically many efficient allocations of a given collection of resources. Therefore, requiring efficiency does not pin down a unique outcome. However, we see later that under certain conditions, efficiency is a good predictor because it then implies sharp restrictions on the choices that can be made. Second, having all the potential benefits of economic activity go to an insatiable, completely selfish person would be efficient because any reallocation of resources would hurt this person and so could not gain unanimous support. Thus, to note that an allocation is efficient is hardly to recommend it on ethical grounds!

However weak efficiency may be as a predictor or as an ethical criterion, actually achieving efficient allocations is extremely demanding. At a macro level, not only must all goods be produced at the lowest possible cost, but the right mix of outputs must be forthcoming, the right levels of savings and investment must be provided, and, in general, there must be no way to increase consumer satisfaction by any reallocation of society's resources. The task of computing an efficient allocation for a complex modern economy is clearly beyond the limits of feasibility. Even with only two people, it may be impossible to determine whether there is any way of rearranging their activities to help one without hurting the other.

Moreover, even if an efficient allocation has been identified, it also is necessary to ensure that the people involved do their parts in bringing it about. The problem is that there often will be inefficient allocations that are better for one person or subgroup than is the target efficient allocation, and these people may be able to effect the inefficient outcome that they prefer.[3] Despite this difficulty, efficiency is both an important device for organizing ideas and a useful criterion for evaluating performance.

### Efficiency of Organizations

Efficiency of outcomes or allocations is not, however, the key concept for the study of organizations. Instead, we are concerned about the efficiency of the organizations themselves. We assume that the fundamental objects that people care about are the

[2] Vilfredo Pareto was an Italian economist and sociologist who is credited with developing this criterion for problems with multiple objectives.

[3] There will necessarily be a third allocation that is efficient and that Pareto dominates the inefficient one: otherwise, the latter would actually be efficient. However, as we will see, identifying and reaching this efficient allocation may be a problem.

outcomes the organizations generate, and that organizations are to be judged on the basis of these outcomes. There are, however, several ways to do this.

A simple way is to make the comparison outcome by outcome. Consider two contracts, routines, decision processes, organizations, or economic systems, say X and Y, which both might be used in a variety of circumstances. Suppose that, in each such circumstance, Y always yields outcomes that are viewed by all the people involved as being at least as good as those that X produces, and that sometimes Y yields results that at least one person definitely prefers to the outcome under X. In this case, X is inefficient because Y does better. In contrast, a contract, routine, process, organization, or system is efficient in this sense if there is no alternative that consistently yields unanimously preferred results. In particular, an organization that always yields efficient outcomes is itself efficient.

This outcome-by-outcome comparison of organizations is quite demanding: To declare an organization inefficient, there must be another that would do better in every possible circumstance. Consequently, the resulting notion of efficiency is weak because it is easy to pass the test of not having a better alternative. Thus, there might be many efficient organizations using this criterion. Later we refine the efficiency notion, for example, by specifying that an organization is inefficient if there is another that does better for each person on average across the circumstances in which the organization operates. Such specifications narrow the class of organizations that meet the test of efficiency. Because it is now easier to find another mechanism that is unanimously preferred, it becomes more difficult to pass the test of there not being a preferred mechanism.

Regardless of which specification we employ, if achieving efficiency of allocations is problematic, realizing the systemic efficiency of organizations clearly is even more demanding. However, the notion of efficiency still provides a key organizing principle.

## Efficiency as a Positive Principle

So far, we have emphasized efficiency as a normative concept, a criterion by which group decisions—including both resource allocation decisions and organization structure decisions—are to be evaluated. However, the quotation at the beginning of this chapter points in another direction. If people seek efficiency in their activities and in the ways they arrange their affairs, then efficiency can become a positive concept, with explanatory and predictive power, as well as a normative, prescriptive one.

There is good reason to expect that people will seek out and settle upon efficient choices. After all, by the definition of efficiency, if an inefficient situation is reached, then someone could propose an alternative that everyone would prefer. If the parties can bargain together effectively and can effectively implement and enforce any agreements they reach, they should be able to realize these gains. Inefficient decisions, whether about resource allocations or organizational arrangements, are thus always vulnerable to being overturned. Efficient arrangements are much less vulnerable because any proposal to change an efficient arrangement will always be opposed by someone. For this reason, we expect to find inefficient arrangements being supplanted over time, while efficient ones survive. We summarize this argument in the following efficiency principle.

icient - survive
icient - supplanted

**The Efficiency Principle:** If people are able to bargain together effectively and can effectively implement and enforce their decisions, then the outcomes of economic activity will tend to be efficient (at least for the parties to the bargain).

Much of our analysis of organizations is based on the efficiency principle. We try to understand existing arrangements as efficient choices, and we interpret changes in these arrangements as efficiency-enhancing responses to changes in the environment within which the arrangements exist.

In pursuing this agenda, it is important to keep in mind the qualifications that bargaining, implementation, and enforcement should be effective. A major focus of the analysis in this book is the task of giving specific meaning to the notions of effective bargaining, implementation, and enforcement, and of identifying and analyzing conditions under which these premises might or might not be expected to hold. What factors influence the possibility and likelihood of efficient bargaining? What factors promote or impede implementation of plans and enforcement of bargains? The answers to these questions provide much of the basis for understanding actual arrangements and for analyzing possible changes.

In using efficiency as a positive concept for predictive purposes, it is especially important to be clear about the set of individuals whose interests are being taken into account in determining economic arrangements. (Recall that efficiency is always defined relative to a specific set of individuals and options.) Suppose a large group of people might be affected by some choice, but only a relatively small subgroup of them are able to communicate with one another effectively, and that they alone can reach agreements and implement and enforce them. Then the appropriate concept for predicting what arrangements will be chosen is efficiency relative to the small, effective group. The interests of people who do not take part in the decision about what choices are made are unlikely to be fully reflected in the choices. For normative, valuative purposes, efficiency relative to the larger group may still be an appropriate criterion, but it is unlikely to be predictively powerful.

efficiency for sub grupo
decided in grupo

Influence cost
or
control (Board)

## The Tasks of Coordination and Motivation

A fundamental observation about the economic world is that people can produce more if they cooperate, specializing in their productive activities and then transacting with one another to acquire the actual goods and services they desire. The problem of organization then arises because when people are specialized producers who need to trade, their decisions and actions need to be coordinated to achieve these gains of cooperation, and the people must be motivated to carry out their parts of the cooperative activity. Both the existence of formal organizations and the specific details of their structures, policies, and procedures reflect attempts to achieve efficiency in coordination and motivation.

People can produce mo
if they cooperate, spe
~~more~~ in their productiv
activity.

Problem of Organi

### Specialization → Need coordination

Adam Smith's famous example of the pin factory vividly shows the benefits of cooperation and specialization and the corresponding need for coordination. Smith described how in his time (the late eighteenth century) the various stages of pin manufacturing were carried out by different people, each of whom specialized in a single task—pulling the wire, straightening it, cutting it to appropriate lengths, sharpening the point, attaching the head, and packaging the finished product—and how the resulting volume of output was many times greater than it would have been if each person involved had done all the stages alone. The crucial point, however, is that such specialization requires coordination. A single person producing pins alone turns out something useful. The time and efforts of the specialists are wasted unless they can be sure that both the people at each of the preceding stages are doing their parts in generating semifinished materials in the appropriate amounts and in a timely way, and that those at the latter stages of manufacturing are prepared to take what the people before have produced and turn it into a finished product.

coordination
Team → View of
Alchian

The principles of specialization and coordination apply both to small, simple economies and to large, complex ones. Robinson Crusoe did not face a coordination problem when he was alone on his desert island, but once Friday entered the story, there were opportunities for gain through specialization and exchange. This meant there was also a need to coordinate the two men's actions to ensure that all the necessary tasks were done and that they were not needlessly repeated by both Crusoe and Friday. In a modern economy, the variety of tasks that are carried out is unfathomably complex. Somehow, each of these jobs must be accomplished, in the appropriate amounts, using the appropriate methods, at the right time, in the right order, and by the right numbers of the right people. Coordinating the billions of people and their choices among the infinite possibilities facing them is a mind-boggling problem. Moreover, even if a solution were somehow found that was reasonably well adapted to currently prevailing conditions, the problem would not be solved once and for all. Conditions at every imaginable level and on every possible dimension are constantly changing, and adaptation to these changes is necessary.

## The Need for Information

A key problem in achieving effective coordination and adaptation is that the information needed to determine the best use of resources and the appropriate adaptations is not freely available to everyone. Efficient choice requires information about individual tastes, technological opportunities, and resource availabilities. No single person in society has all—or even a significant fraction—of the needed information. Instead, information is localized and dispersed throughout the economy. Even if the relevant information were generally available, determining what should be produced, for whom, by whom, and using what methods and materials is an overwhelmingly large and complex problem. Because this information is localized and dispersed, however, no one has the knowledge needed to make these calculations, even if they might be feasible in principle.

Two solutions are possible. Either the dispersed information must be transmitted to a central computer or planner who is expected to solve the resource-allocation problem, or else a more decentralized system must be developed that involves less information transmission and, correspondingly, leaves at least some of the calculations and decisions about economic activity to those with whom the relevant information resides. The trick with the first option is to make timely decisions while keeping the costs of communication and computation from absorbing all the available resources. The challenge of decentralization is to ensure that the separately made decisions yield a coherent, coordinated result.

## Organizational Methods for Achieving Coordination

Different organizational structures achieve coordination in different ways and with differing results. As we discuss in Chapter 1, the original strategy of the Hudson's Bay Company was overly centralized. As a result, decisions were not timely and the company made poor use of local information. In contrast, the structure of General Motors was initially overly decentralized, and the company suffered from a consequent lack of coordination. In the case of Smith's pin factory, the solution was a single firm whose owner-manager specialized in providing coordination. He or she hired workers and assigned them to the different tasks, set the levels at which each was to perform, tracked performance and the external environment, and adjusted plans as needed. This individual also probably owned the capital equipment that was being used, collected the sales revenues, and paid the bills. Another solution might have been a cooperative of pinmakers, where the workers would jointly have decided on the levels of activity and the task definitions and assignments and would then have shared in the costs and revenues they generated. The older, highly decentralized system of individual craftsmen, each producing pins alone, sacrificed the gains from specialization but reduced the need for coordination. Yet another solution would have been to organize each stage of production as a separate firm and let the transactions between stages be intermediated by the market. This last alternative may sound far-fetched if we picture the person at each stage in the pin factory as a separate firm, buying input from the person/firm on one side and selling output to the next person/firm in the production line. However, the striking differences between GM and Toyota in their reliance on independent suppliers (see Chapter 1) indicate that this kind of alternative is genuine. Finally, within some modern corporations, products are sold by one division to another using *transfer prices*, and the division managers are judged on their individual division's profitability. With this system, the firm's internal organization mimics the market in many ways.

COORDINATION THROUGH A SYSTEM OF MARKETS AND PRICES A thoroughgoing use of the market is one possible solution to the problem of coordinating economic activity. At the extreme, all transactions could be between separate individuals on an arm's-length basis, and there would be no firms or other organizations apart from the market system itself. The opposite extreme would be complete elimination of the price system under a regime of explicit central planning, with all decisions being made within a single (presumably multilevel) organization. Of course, no economic system approaches either extreme. Even in their most centralized versions, the centrally planned communist economies left many decisions to individual consumers who made their choices in part in response to prices. The market economies feature firms that interact with one another through markets but within which activity is explicitly coordinated through plans and hierarchical structures.

In fact, the system of markets and prices is often a remarkably effective mechanism for achieving coordination. Day in and day out, without any conscious central direction, it induces people to employ their talents and resources so effectively that the shortages and rationing which are familiar to residents of planned economies are deemed newsworthy events when they occur in market economies. As a practical matter, the advantages of the market system over socialist planned economies seem clear.

As we see in the next chapter, it is even possible to argue formally that no system can solve the coordination problem more effectively than a system of markets coordinated by prices. A mathematical model is used to show that in economies with certain characteristics, the allocations generated by a price system are always efficient for society as a whole. Moreover, under certain conditions, the price system achieves this result while economizing on information demands—the system requires transmitting less information than any other system capable of ensuring efficient outcomes (see Chapter 4). In an ideally functioning system of markets, all that anyone needs to know is his or her own capabilities and tastes and the prevailing prices. There is no need to transmit detailed information about preferences, technological possibilities, resource availabilities, and the like that would be needed to achieve a centralized solution because the prices summarize all the relevant information. Furthermore, when conditions change, detailed local knowledge of these changes need not be transmitted to achieve effective responses. Instead, the changes in prices again convey all the information that is actually needed for people to respond effectively.

Information is so important.

INCENTIVES IN MARKETS The strength of a market system with private property lies not just in its providing the information needed to compute an efficient allocation of

nmation +
centive
important to
hieve the
ficiency

ve depends on
lf interest.

sfer ownership rights
ht affect the
avior.

efficient manner, however. At least equally important is the manner in which it accepts individually self-interested behavior, but then channels this behavior in desired directions. People do not have to be cajoled, artificially induced, or forced to do their parts in a well-functioning market system. Instead, they are simply left to pursue their own objectives as they see fit. Yet, at least in the right circumstances (which we explore in Chapter 3), people are led by Adam Smith's "invisible hand" of impersonal market forces to take the actions needed to achieve an efficient, coordinated pattern of choices. Workers, selfishly attempting to maximize their own individual welfare, are led to select the training, careers, and jobs where their talents and energy are most valuable. Producers, pursuing only private profits, are led to develop the goods and services on which consumers put the highest values and to produce these goods and services at the lowest possible costs. The owners of resources and capital assets, seeking only to increase their own wealth, are led to deploy these assets in socially desirable ways. Finally, consumers, seeking only to satisfy their individual wants and needs, are led to do so in the way that puts the least strain on society's resources for the level of satisfaction achieved.

All of this is based on a particular theory of markets, which posits that competition is ubiquitous, firms have little market power, and the only goods that are of significance are those that are traded in active markets. The incentives provided by real markets do not always align so nicely with the realization of social objectives. Large firms or cartels may set prices inefficiently high, leading to inefficient resource allocations. Externalities and missing markets for some goods may lead to additional distortions. The quality of goods may be hard to verify, leading some consumers to make mistakes in their choice of goods and some firms to neglect quality control in the hopes that their actions will go unnoticed. As we see in the examples of Chapter 1, similar failings plague other forms of organization as well. Organizations either must rely on individuals to ignore their own self-interests, with unsurprisingly disappointing results, or else they must devote ingenuity and resources to bring coherence between individual self-interest and the social or organizational objectives.

For example, Salomon Brothers' complex system of attributing profits and paying bonuses linked to performance is an attempt to generate within the firm the sort of incentives that are provided automatically by the market. The transferring of partial ownership to the employees is intended to give them the incentives that owners have in order to care about the long-term value of their asset, the firm. The manager and workers in the Soviet factory that produced the single gigantic nail were responding to inappropriate incentives. Their jobs were made easiest by adopting a socially wasteful production plan, and their incentives were to minimize their efforts while meeting their poorly specified quota. Such a gross inefficiency would never arise in a market system, although, as we see in later chapters, more subtle difficulties can be expected.

## Transaction Costs Analysis

If markets can perform so well, why then do we so often see the price system supplanted, with economic activity being organized within and among formal hierarchical structures using explicit planning and directives? More simply, why are there firms? What is their economic function? And what determines which transactions are mediated through markets and which are brought within a formal organization and made under centralized direction?

These fundamental questions were first posed by Ronald Coase. According to Coase, there are costs to carrying out transactions, and these **transaction costs** differ depending on both the nature of the transaction and on the way that it is organized. Furthermore, as suggested by the efficiency principle, the tendency is to adopt the organizational mode that best economizes on these transaction costs. Thus, transactions tend to occur in the market when doing so is most efficient, and they are brought within the firm or some other formal organization when doing so minimizes the costs of carrying them out.

This is a simple, but profound, idea. However, Coase was not very explicit about the origin and nature of these transaction costs, and without a systematic understanding of these issues, the idea is not very useful. Consequently, much of the research in the economics of organization has been devoted to giving substance and content to his idea. In fact, transaction costs are the costs of running the system: the costs of coordinating and of motivating. Thus, under the hypothesis that organizational structure and design are determined by minimizing transaction costs, both aspects of the organization problem affect the allocation of activity among organizational forms.

### Types of Transaction Costs

{coordination costs
motivating costs

Different organizational forms and institutional and contractual arrangements represent different solutions to the problems of coordination and motivation. These problems give rise to transaction costs, which manifest themselves differently in different contexts.

**Coordination Costs** Under a market system, transaction costs associated with the coordination problem arise from the need to determine prices and other details of the transaction, to make the existence and location of potential buyers and sellers known to one another, and to bring the buyers and sellers together to transact.

As an example of these coordination costs, think about the problem of exchanging financial assets such as stocks and bonds. These transactions mostly take place through organized financial exchanges, like the New York, London, and Tokyo stock exchanges. Very few markets function more efficiently than the organized financial markets, and yet the amount of resources absorbed in their operation is clearly significant. Large buildings, immense communication and computational power, and the talents of thousands of gifted people are employed in setting prices and carrying out transactions. If the often breathtaking incomes of investment bankers and security dealers are any indication of the social costs involved in employing them in this industry, the transaction costs of running these markets are very large.

In other markets, transaction costs associated with coordination include the resources that sellers expend on market research to determine buyers' tastes, on advertising and marketing expenditures to make the product or service known, and on managerial decisions determining the prices to charge. On the buyers' side, they also include the time spent searching for suppliers and for the best prices. More subtly, the transaction costs also include the lost benefits that are not realized because the matching of buyers and sellers is imperfect and worthwhile transactions fail to occur.

coordination cost [se
ex: resources to ~~do~~ m
the buyer's tastes,
advertising market
managerial deci

The transaction costs of coordination through hierarchies—whether private or governmental—are primarily the costs of transmitting up through the hierarchy the initially dispersed information that is needed to determine an efficient plan, using the information to determine the plan to be implemented, and then communicating the plan to those responsible for implementing it. These costs include not only the direct costs of compiling and transmitting information, but also the time costs of delay while the communication is taking place and while the center is determining the plan. Because this communication can never be perfect, there are also transaction costs of maladaptation that occur because decision makers have only insufficient or inaccurate information.

**Motivation Costs** The transaction costs associated with the motivation problem are primarily of two kinds. The first type of costs are those associated with informational

**incompleteness and asymmetries**—situations in which the parties to a potential or actual transaction do not have all the relevant information needed to determine whether the terms of an agreement are mutually acceptable and whether these terms are actually being met. For example, the potential buyer of a new car may have difficulty determining whether the seller's claims about its economy and reliability are correct, and may wonder why the seller would want to get rid of the vehicle. Or a sales manager may have difficulty in determining whether a salesperson in the field is actually devoting full time and honest effort to the company's business, or instead is pursuing private interests on company time. In such circumstances, mutually advantageous transactions may fail to occur, because one or the other party fears being victimized, or costly arrangements will be made to protect against opportunistic behavior.

The second type of transaction costs connected to the motivation problem arise from **imperfect commitment**—the inability of parties to bind themselves to follow through on threats and promises that they would like to make but which, having made, they would later like to renounce. As an example, consider a manufacturer seeking to have a supplier make a large investment to meet the manufacturer's specific needs. The supplier must be concerned that—all promises to the contrary not withstanding—once the investment is sunk the manufacturer will try to force a lower price and other concessions on the supplier, who will then have little recourse. Recognition that threats and promises may not be kept deprives them of their credibility. Thus, far-sighted people will not rely upon them, and again there will be missed opportunities or a necessity of expending resources to facilitate commitment or protect against opportunism. The manufacturer would gain if it were possible to commit not to behave opportunistically, because then the supplier would be more willing to make the investment. Achieving such commitment may be difficult, and so the investment may not be made or costly measures might need to be put in place to defend the supplier.

These problems affect both market and nonmarket organizations, although their nature and impact may differ between organizational forms. Therefore, one form may be better adapted than another for a specific transaction.

## Dimensions of Transactions

According to the transaction costs approach, the variety of ways of organizing transactions found in the world reflects the fact that transactions differ in some basic attributes. Five kinds of transaction attributes play important roles in our analysis:

1. the **specificity** of the investments required to conduct the transaction
2. the **frequency** with which similar transactions occur and the **duration** or period of time over which they are repeated
3. the **complexity** of the transaction and the **uncertainty** about what performance will be required
4. the **difficulty of measuring performance** in the transaction
5. the **connectedness** of the transaction to other transactions involving other people

uncertainty

ASSET SPECIFICITY One important dimension on which transactions differ is the nature of the investments that the parties to a transaction must make. When an individual consumer buys bread from a baker, neither party makes any investment with that particular transaction in mind. The baker may invest in a store and an oven, but he or she uses those assets to supply many different customers. In contrast, when a subcontractor makes wing assemblies for a particular model of Boeing airliner, it may invest in setting up a production line to make those specific assemblies. Such an investment is called a **specific investment** because it would lose much of its value outside of the specific use of providing wings to Boeing. The subcontractor would not want to make the investment unless it has a firm order from its customer, or at least reasonable assurance that an order will be forthcoming. For the same reason, an employee may not want to invest in learning the systems of a declining company where the prospects of continuing employment are poor. Transactions that require specific investments normally also require a contract or practice to protect the investor against early termination or opportunistic renegotiation of the terms of the production relationship.

FREQUENCY AND DURATION Some transactions are one-time affairs as, for example, when a homeowner buys a house from its previous owner. Others are repeated frequently, involving some of the same parties under more or less similar conditions over a long period of time.

In the first case, one expects the parties to use whatever general mechanisms are available in the community to control their transaction. In particular, they likely will resort to a standard form contract, with any disputes between them to be resolved in court.

In the case of parties who interact frequently, one expects quite a different sort of mechanism that is specifically geared to the particulars of their relationship. For example, disputes between a supervisor and worker in a factory are rarely resolved in courtrooms. Instead, factories may set up a special grievances committee involving the union or other worker representatives, or an ombudsman may be used to hear complaints and attempt to mediate a solution. The special purpose institution is worthwhile because it can be tailored to the particular circumstances of the factory, keep down the cost of resolving disputes, and be continually improved in light of the circumstances in the particular factory. Generally, when similar transactions occur frequently over a long period of time involving some of the same parties, the one who interacts repeatedly may find it valuable to design and introduce low-cost routines to manage the transaction.

Frequency and duration also have another effect. Parties involved in a long, close relationship with frequent interactions have many opportunities to grant or withhold favors to one another. The ability to reward faithful partners and punish unfaithful ones in a long-term relationship greatly reduces the need for any kind of formal mechanism to enforce agreements between them. The parties can also develop understandings and routines that reduce the need for explicit planning to coordinate their actions. These practices can sometimes eliminate the need for formal, detailed agreements, both because the parties understand what is expected of them and because they have no need to document those understandings for outsiders to enforce. The cost savings that result can be considerable.

UNCERTAINTY AND COMPLEXITY The standard way for two parties to organize a market transaction is to write a contract specifying what is expected from each. If the product is wheat, then the contract may simply specify that a fixed amount of a standardized grade of the grain (for example, Manitoba #1 Northern hard wheat) will be delivered at a particular date (say, April 1, 1992) and place (say, Winnipeg) for a specified price (for example, $5 Canadian per bushel). The basic contract is then very simple.

In contrast, a contract to build a power plant for an electric utility is very complicated. The utility's estimate of demand may change during construction, and the cost and availability of different kinds of fuel may change as well. The environmental impact of the facility may be unknown at the outset, and the cost of providing the necessary environmental and health and safety safeguards may be unpredictable. The right way to proceed, the length of time to take, and whether to finish the project at all are decisions that will have to be made later, after the contract is signed and execution begins. If the project is changed or delayed or terminated, there will need to be some way to determine what payments ought to be made.

Uncertainty about the conditions that will prevail when a contract is being executed together with complexity of the task make it impossible, or at least uneconomical, to determine in advance what should be done in every possible contingency, so the contract that is written will generally be less determinate than in a simpler setting. Rather than specifying how much of what is to be delivered when, the contract may specify who has the right to make which decisions and within what limits.

Returning to our example of a maker of wing assemblies for an aircraft, the contract between the aircraft manufacturer and the supplier might deal with uncertainty about future aircraft sales by specifying that the supplier will provide whatever number of assemblies the buyer requires according to a particular pricing formula. In return, the buyer may promise in the contract to purchase wing assemblies only from that supplier as long as the supplier is able to meet the demand. The buyer might also promise to supply advance estimates of demand that are within a fixed percentage of the actual quantities, to subsidize the purchase of specific assets used in the production, and so on.

Generally, when uncertainty and complexity make it hard to predict what performance will be desirable, contracting becomes more complex, specifying rights, obligations, and procedures rather than actual performance standards.

**Difficulty of Performance Measurement** Even when the desired performance is perfectly predictable, it may be difficult or costly to measure actual performance. For example, a person who employs a lawyer in a divorce proceeding may have no idea whether the negotiated settlement is really a good one, or whether a better lawyer could have negotiated a better deal. Similarly, the low output of a group of workers in a factory might be due to low effort, to poor materials, or to inferior methods used by the company. When a taxi that has been driven by several drivers over a period of time breaks down, the owner may be unable to tell which (if any) of the drivers has abused the car or failed to get maintenance when needed, or whether instead the breakdown is due to poor design or plain bad luck. Of course, if the taxi doesn't break down immediately, but hard use now makes future problems more likely, the cost of that abuse is nearly impossible to measure.

As these examples suggest, it is hard to provide effective incentives unless one can measure performance accurately. If the lawyer's performance could be accurately evaluated, or the factory workers', or the taxi drivers', then the consumer, factory manager, or taxi company owner could hold those parties responsible for their performance. That would presumably lead to more effort and better results.

When measuring performance is difficult, people commonly arrange their affairs to make measurement easier or to reduce the importance of accurate measurements. In our taxi example, the taxi may be assigned to just one driver, so that responsibility for any evident damage can be more easily assigned. Or, the taxi may be driver owned, so that any loss from abuse or poor maintenance (including even losses that are not immediately detectable) comes straight out of the owner/driver's pocket. Other attributes of the transaction determine which of these possible solutions is best, or even whether any of them are workable.

**Connectedness to Other Transactions** Finally, transactions differ in how they are connected to other transactions, especially those involving other people. Some transactions are largely independent of all others. For example, an office's decisions about when to buy new typewriters, where to keep its files, and which supplier to use for general office supplies hardly need to be coordinated.

Other transactions are much more interdependent. When railroads were introduced in the United States in the nineteenth century, the various railroad companies failed to coordinate their choices of track gauges (the size of the rails and distance between them). Because rail cars adapted to one gauge of track cannot be used on track laid to other gauges, the result was that goods being shipped long distances had to be unloaded and reloaded onto different cars at several points in the journey. Standardization on any one of the various gauges that were actually adopted would have been much more efficient. A similarly costly situation is still present in Europe, where Spanish rail gauges do not match those used in France. The eventual standardization of the rail gauges in the United States resulted in much quicker and less expensive shipping of goods and contributed to the development of the western parts of the country.

For a more modern example, suppose a computer maker is designing a new model of computer. It cannot deliver a working model until all the relevant components such as the memory chips, central processor, power supply, and so on are all available in sufficient quantity to begin assembly. In addition, the operating system for the computer and some of the application software must be ready; otherwise, there is nothing useful that the finished computer can actually do.

The manufacturers of the various parts and the software developers in this situation need to have their activities closely coordinated. There is little advantage to rushing the completion of a plant to assemble the computer, for example, if the other parts of the system will not be ready to go. Similarly, the amounts of the various components that the different suppliers are able to deliver must line up with one another, for there is little value to having a larger number of keyboards than disk drives. The capacities and capabilities of the machine components must be compatible with one another, and the design tolerances need to be coordinated. Failures to align capacities or to match design tolerances or to have components ready on time can be much costlier than failures to adopt the best possible design and introduction date.

When the costliest potential mistakes are of these kinds—rather than, say, failures to make effective use of local resources—we say that the transactions display **design connectedness**. Design connectedness is just one extreme; the relative costs of different kinds of mistakes can in general have any relationship.

One way that firms respond to close connectedness is to strengthen central coordination mechanisms. This may mean that there are more meetings among the people involved in the individual transactions, or that the managers in charge spend more time on oversight, or some combination of those things. A second way firms respond is to reduce the number of different people involved, so that fewer people need to be coordinated. The particular way that close connectedness is managed depends on other attributes of the connected transactions.

## Limits of the Transaction Costs Approach

This transaction costs approach is appealing, and we adopt it later for some of our analyses. However, the approach cannot be correctly applied to all problems in economic organization because, without additional conditions, its fundamental argument—that economic activity and organizations are arranged so as to minimize transaction costs—is problematic. There are two main problems.

First, it is not generally true that the total costs of an economic activity can be expressed as the sum of production costs and transaction costs, where the former depend only on the technology and the inputs used and the latter depend only on the way transactions are organized. Production and transaction costs generally depend both on the organization and on the technology, which makes the conceptual separation between production and transaction costs troublesome. If production is lost due to delays in planning, is it the result of slow planning or of a technology that cannot adapt quickly to late changes in the plan? A more subtle example can be seen in the semiconductor industry.[4] Integrated circuit production is marked by increasing returns to scale and very strong learning curve effects, so that costs are lower the larger the volume of production within any facility, both in any single time period and in aggregate over time. Thus, efficiency in production would require that any particular design of a circuit be produced by a single manufacturer. For a long time, however, it was standard procedure for a company that developed a new chip to give the design to a second firm that would compete with it in producing and selling the chip. It was even common to assist this "second source" in setting up production.

This way of organizing the production of integrated circuits sacrifices production cost efficiencies for other advantages: Without a competing second source, potential users of the new integrated circuit would be reluctant to adopt it for fear that once they had become locked into its use, the supplier would exploit its monopoly position. Creating a competing second source is an effective way to achieve commitment, leading to increased demand.[5] Are the extra costs incurred a "production cost," arising because an inefficient technology is used that does not take full advantage of economies of scale, or a "transaction cost," incurred to satisfy customers that the terms of the transaction are secure? There are no unassailable answers to these questions. The lesson is that although the costs of transacting are real, they are not always easily separated from other kinds of costs.

The second problem is not with the concept of transaction costs per se, but with the notion that efficient institutions would minimize them. For example, according to Coase's postulate, employment relations can be understood as minimizing the total transaction costs involved. But why should employers minimize total transaction costs in designing their employment, compensation, promotion, supervision, and performance review systems, rather than simply the categories of costs that they themselves must bear? Some of the transaction costs surely will be borne by the employees; why should we expect the employers' choices to take proper account of these? In fact, why would they not push all the transaction costs onto the employees?

A standard answer to these questions is that competition would force employers to take account of the costs to employees. In Chapter 8, we argue that this standard answer is of limited application. But even when it does apply, having to rely on competition or other external forces to bring about efficiency would critically weaken the theory because its range of potential application would then be severely narrowed.

A more general version of this second problem is that because there are typically many quite different efficient solutions to any resource-allocation problem, efficiency alone may not be a strong enough criterion to give very specific predictions or clear explanations. Too many different patterns of organization might be compatible with efficiency for it to be a useful concept.

[4] See Andrea Shepard, "Licensing to Enhance the Demand for New Products," *Rand Journal of Economics*, 18 (1987), 360–68, and Joseph Farrell and Nancy Gallini, "Second-Sourcing as Commitment: Monopoly Incentives to Attract Competition," *Quarterly Journal of Economics*, 103 (1988), 673–94.

[5] The break with this practice came when Intel chose not to second-source its 80386 microprocessor. The market for this chip was judged to be secure even without second-sourcing because Intel's performance was secured by its need to compete with its own earlier chip designs, which continued to be produced by other manufacturers.

It turns out that one simplifying assumption, that is, the condition of no wealth effects discussed in the next section, takes care of this problem completely.[6] When this condition is satisfied, only one pattern of behavior is consistent with efficiency, and that is the pattern that maximizes the total value created in the transaction.

## Wealth Effects, Value Maximization, and the Coase Theorem

In many economic decisions, the choice actually made depends on the decision maker's wealth. A poor person (or a poor country) may not have the resources to pursue some courses of action that a richer one could. Even when the same alternatives are affordable, a poorer person might still make different trade-offs than a richer one. For example, the poor person might be reluctant to take financial risks that a rich person would welcome. The changes in choices resulting from increased wealth are known as **wealth effects.**

### The Value Maximization Principle

Although wealth effects can sometimes be significant, this is not always the case. In fact, the formal analysis of problems in the economics of organizations is greatly simplified when wealth effects can be ignored entirely. Moreover, it is precisely in ignoring wealth effects that such key management concepts as "creating value" become unambiguously defined.

No Wealth Effects We say that there are *no wealth effects* for a certain decision maker with respect to a set of possible decisions when three conditions hold. First, given any two alternative decisions $y_1$ and $y_2$, there is a definite amount of money $\$x$ that would be sufficient to compensate the decision maker for switching from $y_1$ to $y_2$ (or from $y_2$ to $y_1$). Second, if the decision maker were first given an additional amount of wealth, then the amount needed to compensate the decision maker for the switch from $y_1$ to $y_2$ would be unaffected. Third, the decision maker must have enough money to be able to absorb any wealth reduction necessary to pay for a switch from the less preferred to the more preferred option.

conditions to no wealth effects = [illegible]

None of these conditions can be expected always to hold. For some people, for example, there may be no amount of money that they would accept as compensation for a serious risk of loss of life or limb, or for being forced to live far from family and their childhood homes, or to live in a culture where they cannot exercise their religious beliefs. Nevertheless, the condition that there is some monetary amount that would compensate for a change of circumstances holds widely for many of the most common kinds of business decisions.

first condition

To examine the implications and applicability of the second condition, suppose that a corporation suddenly finds itself richer on account of an unexpected increase in the value of its assets. If there are no wealth effects, then the price the corporation would demand for its goods and the returns it would demand from its planned investments would remain unchanged. In this example, the absence of wealth effects

second condition

[6] Like other modelers, economists use the terms *assumption* and *assume* in a different sense than they are often used in ordinary discourse. In everyday conversation, assuming something connotes believing it is true. Making an assumption in an economic model carries no such connotation. An assumption is merely a working hypothesis used to abstract from the complexity of the real economic world. The purpose of making an assumption may be to derive a good approximate prediction or to highlight a single force or effect for closer study and better understanding. Assumptions are used in both ways in this text.

seems likely to hold, at least over a broad range of wealth levels. As a second example, however, suppose a worker wins the jackpot in the state lottery. If there are no wealth effects with respect to current consumption choices, then the winner would not buy a new house, or quit his or her job, or do any of the things lottery winners normally do. Here the assumption of no wealth effects seems particularly inappropriate.

The third condition connects the decision maker's initial wealth and the changes in the nonfinancial situation that are being considered. The amount by which a worker's pay could be cut to offset his or her psychic gain in being allowed to shift working hours to miss rush-hour traffic is probably small relative to income, and the condition is likely to hold in this case. On the other hand, holding a worker in a nuclear reactor facility financially responsible for the effects of his or her mistakes is apt to run into wealth constraints, and so the assumption of no wealth effects will be inappropriate.

In general, these examples suggest that the assumption of no wealth effects is most restrictive—least likely to be valid and most likely to lead to incorrect conclusions—when the decision makers are individuals and when large cash transfers or significant changes in personal living conditions are involved. When the sizes of the cash transfers are small relative to the decision maker's financial resources, assuming that there are no wealth effects (or that they are small enough that they can safely be ignored) is more likely to be a good approximation to reality.

equivalent value

equivalent value

THE EQUIVALENT VALUE INDEX The utility function of a decision maker who shows no wealth effects with respect to a set of decisions can be represented very simply.

Let $x$ represent the decision maker's monetary wealth and let $y$ be a list of all the other influences or characteristics associated with decisions that affect his or her preferences: social approval, job assignment, effort exerted on the job, and so on. An important example involves uncertain income and expenses, where $x$ is interpreted as the certain, unconditional amount of money that will be received (or the average, expected amount) and $y$ reflects a risky component of income. In general, the utility function takes the form $u(x,y)$, where $x$ and $y$ may interact in complex ways. If there are no wealth effects, however, then there is always a cash equivalent value $v(y)$ that can be assigned to the list $y$ and the decision maker's utility function can be written in the form $u(x,y) = x + v(y)$. In other words, by adding the cash equivalent value $v(y)$ to the decision maker's wealth $x$, we obtain an index of personal welfare, which we may call his or her *value index*.[7] The importance of the value index is that, when it is valid, a related index—the total value of the affected parties—is an appropriate measure of welfare changes for group decision making.[8] We state the matter as follows.

> **The Value Maximization Principle**: An allocation among a group of people whose preferences display no wealth effects is efficient only if it maximizes the total value of the affected parties. Moreover, for any inefficient allocation, there exists another (total value maximizing) allocation that *all* of the parties *strictly* prefer.

Important to look if it apply to collective action

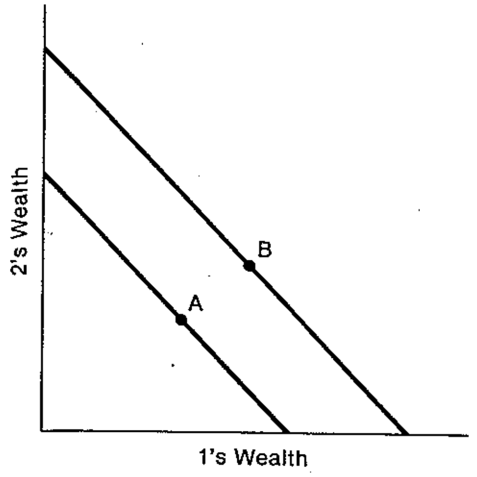

**Figure 2.1:** Any point like A on a line of lower total wealth is Pareto dominated by some point B on the line of highest total wealth. Points like B on the line of highest total wealth are undominated.

## *The Logic of Value Maximization*

To establish the principle for a particular example, consider an investment decision involving two people with utility functions that satisfy the condition of no wealth effects: $U_i(x,y) = x + v_i(y)$, $i=1,2$, where $y$ represents inputs to be provided by the parties.[9] The investment generates total cash income of $P(y)$. We may think of $v_i(y)$ as representing the personal cost investor $i$ bears in supplying the agreed inputs. In that case, $v_i(y)$ is a negative number for positive levels of $y$. The income $P(y)$ will be divided between the individuals, with $x_1$ being paid to individual 1 and $x_2$ being paid to individual 2, so that $x_1 + x_2 = P(y)$. For any particular allocation $(x_1,x_2,y)$, *the total utility or value of the two parties is* $[x_1 + v_1(y)] + [x_2 + v_2(y)]$, which (because $x_1 + x_2 = P(y)$) is equal to $P(y) + v_1(y) + v_2(y)$. *The total value depends on y alone and not on the profit shares x.* As we vary the profit shares $x_1$ and $x_2$, the individual utilities of the two parties change, but the total utility remains fixed.

Figure 2.1 illustrates the situation. Each line depicts the possible levels of value for the two parties for any fixed investment decision $y$, as the profit shares $x$ are varied. The fact that the lines are straight and run at 45° to the axes reflects the fact that total value is independent of its distribution between the parties; it is possible to move utility or value from one party to the other (via the xs) without affecting the totals. It is clear from the diagram that any point like A on a line that corresponds to lower total value is Pareto dominated by some point like B on the line of highest total value. It is also clear from the figure that no point on the line of highest total value is Pareto dominated by any point on that line or any other line. Consequently, *an allocation $(x_1,x_2,y)$ is efficient if and only if y maximizes the total value:* $P(y) + \boxed{v_1(y) + v_2(y)}$ → this is how the value will be created

$\Delta X + V(Y_1) =$

The intuition of this result is simple: With more total it is always possible to distribute it in a way that makes everyone better off. A full mathematical proof of the proposition is developed in the exercises at the end of the chapter.[10]

[7] To see how the three preceding conditions relate to this formula, note first that the change in $x$ necessary to compensate for a change from $y_1$ to $y_2$ is easily computed to be $\Delta x = v(y_2) - v(y_1)$, because for any given initial wealth $M$, $M + \Delta x + v(y_2) = M + v(y_1)$. The change in $x$ to compensate for switching from $y_2$ to $y_1$ is then just $-\Delta x$. The calculated amount $\Delta x$ is independent of the initial wealth, $M$, as the second condition requires. With a general utility function $u(x,y)$, there might be no amount $\Delta$ that would make the equality $u(M+\Delta, y_2) = u(M, y_1)$ hold, as the first condition requires, and even if there were such an amount, its magnitude would generally depend on $M$. Finally, as long as the initial wealth $M$ is larger than the largest possible difference in the $v(y)$ values between two alternative $y$s, so that making the transfer $\Delta x$ cannot require more money than the individual has, then the third condition holds.

[8] The total value is also sometimes called the total (consumer and/or producer) surplus.

[9] We can think of $y$ as being the pair $(y_1,y_2)$, where $y_i$ is $i$'s contribution. Then writing $v_i(y)$ rather than $v_i(y_i)$ allows the possibility that each individual's costs depend on both parties' contributions. At the same time, writing $v_i(y)$ allows that $v_i$ depends only on $y_i$.

[10] In the presence of wealth effects, an allocation that maximizes the sum of utilities is still efficient, but there may also be efficient allocations that do not maximize the sum of utilities.

APPLYING VALUE MAXIMIZATION Although this discussion is set in the context of two people making an investment, the principle itself is much more general. When preferences take the form we have described, then any decision $(x,y)$ is efficient if and only if $y$ is chosen to maximize the total value of the parties. Importantly, the efficiency of the choice $(x,y)$ does not depend on the selection of $x$, which determines only how the fruits of the joint enterprise are shared. When the value maximization principle applies, the issue of the distribution of value is completely separable from the issue of how value is created. While this separation is not always realistic (see Chapter 8), it is often reasonable and it does always simplify the analysis of economic organization problems. For that reason, it is a good starting point for thinking about organizations.

only where there isn't wealth effects

The abstract model we have described has wide application because the variable $y$ can be given so many different interpretations. The model can apply to people in a community who must decide on the resources $y$ to be devoted to parks, libraries, or some other public good. It can also apply to overtime assignments ($y$ is then the identity of the person assigned to work overtime), to water rights ($y$ identifies the owner of the rights), to the brand of computer to be used ($y$ names the brand), or to the location of the new office building ($y$ specifies the location).

pareto?

## The Coase Theorem

os benefícios
são ÷ entre os
só ocorrerá sendo
viável porém que é
haver ganhos mútuos.

In practical business situations, the way the benefits of an agreement are divided among the parties will depend, of course, on what assets each brings to the bargaining table, how patient each party is, what outside opportunities are available, and so on. Nevertheless, if the parties engage in efficient bargaining, that is, if they reach an agreement from which there is no possibility of further mutual gain, and if the value maximization principle applies, then regardless of what cash changes hands at the time of agreement, $y$ will be chosen to maximize the total value of the parties to the agreement. Only the distribution of the costs and benefits will be affected by the strength of their relative bargaining positions, and this distribution will show up in the $x$s. This conclusion is summarized in a proposition that is also due to Coase:

> **The Coase Theorem** If the parties bargain to an efficient agreement (for themselves) and if their preferences display no wealth effects, then the value-creating activities ($y$) that they will agree upon do not depend on the bargaining power of the parties or on what assets each owned when the bargaining began. Rather, efficiency alone determines the activity choice. The other factors can affect only decisions about how the costs and benefits are to be shared ($x$).

This celebrated proposition is the foundation of the transaction costs approach to the theory of the firm and other economic organizations. With the assumption of no wealth effects, the Coase theorem and the efficiency principle mean that all real activities are determined to maximize tne total value of the parties, taking into account the costs of organization (transaction costs) along with all other kinds of costs. For any given plan of production specifying who is to make what with which resources (and thus the aggregate production costs that will be incurred), if we think of transaction costs as the costs of managing the transactions (including the costs of writing contracts, supervising workers, enforcing contracts, and resolving disputes), then the efficient organization for that plan is the one that minimizes transaction costs.

To understand the significance of this perspective, it is helpful to contrast it with other perspectives that have been vigorously advocated.

### The Transaction Costs Approach versus Alternative Views

The transaction costs approach differs sharply from the Marxian approach. According to Marxian theorists, organizational arrangements are a reflection of underlying power relationships and class interests. In contrast, according to the transaction costs perspective, the choice of a firm's organization ($y$) does not depend on the a priori distribution of power between the owners of capital and the laborers. For example, in Yugoslavia, where employee control of firms has been the norm, labor hired managers who organized the factories in similar ways and instituted similar kinds of controls (even over labor) as those seen in capitalist firms.

Applied to relationships among firms and between firms and their customers, the transaction costs approach suggests that business arrangements should be understood as attempts to increase the total wealth available for sharing among the parties. Scholars trained in the Harvard school of industrial organization and antitrust would try instead to explain the arrangements as attempts by the firms to increase their ability to manipulate prices for the products they sell or the inputs or labor they buy. For example, resale price maintenance—the practice by manufacturers of limiting contractually the rights of distributors and retailers to set prices—has been attacked by members of the Harvard school as anticompetitive. The transaction costs approach suggests that it in fact may be efficiency promoting. Adherents of this approach might argue that, without resale price maintenance, discounters would free-ride off the service and customer education provided by other dealers, relying on them to provide product support but then grabbing the sales for themselves. This would drive the amount of such support below efficient levels.

The correct way to decide between competing hypotheses is to confront them systematically with detailed evidence. The observation concerning management methods in the Yugoslav firm represents an instance of this, but it is by itself insufficient. We report more of the relevant evidence later in this book.

In any case, in application, it is important to remember that the Coase theorem and its various implications depend on restrictive hypotheses regarding preferences and, perhaps more importantly, on the ability to make unlimited transfer payments between the parties. The implications do not hold when some of the parties have very limited capital with which to make payments. Thus, although it would be reasonable to apply this analysis to study the terms of a contract between General Motors and Toyota, it would be a mistake to apply it uncritically, for example, to study land tenure in a developing country or the institution of slavery in the pre-Civil War American South.

## ORGANIZATIONAL OBJECTIVES

It is traditional in economics and management texts to assume that firms seek to maximize profits and, more generally, to ascribe well-defined goals to organizations and to presume that the organization acts in pursuit of these objectives. Occasionally we employ such a hypothesis in this text, when it is convenient to treat the organization as a purposeful entity. More often, however, we do *not* presume that organizations per se have goals that they seek to realize. Rather, we are careful to treat organizational decisions and actions as the outcomes either of strategic interplay among self-interested people responding to incentives designed to influence their behavior, or of collective or managerial attempts to compromise the interests of the parties affected by the decisions. Only when the value maximization principle applies is there an objective that we can ascribe to the firm that is implied by considerations of efficiency alone.

## Profit Maximization

The goal most commonly ascribed to firms in economic analyses is profit maximization. It might seem at least that the self-interested owners of a firm would unanimously favor such a goal. In that case, they would attempt to design the organization to motivate its managers and employees to pursue profits. In fact, we often conduct our analyses presuming that this is the case, but the reader should remember that there are many reasons why owners might have other objectives.

coops

First, to the extent that one of the owners is also a customer of the firm or one of its input suppliers, that owner might prefer that the firm not maximize profits in dealing with him or her, but instead favor the owner with better prices or terms. This would be a potential problem in a firm where ownership is shared between employees and outside investors, with each owning part of the claims on the firm. The inside employee-owners might prefer that the firm's managers adopt policies to protect workers' jobs, pay high wages, and provide many on-the-job benefits. Meanwhile, the outside investor-owners might prefer that the value of their investment in the firm be maximized.

Problem with horizonal among investors

Second, many of the decisions of the firm involve expenditures and receipts that are both uncertain and spread out over time. In such cases, it is common to assume that people are interested in the expected value of the discounted stream of utility they receive over time in the various uncertain future circumstances (see Chapter 12). However, just as owners differ in the likelihoods they place on the various ways future events might unfold and in the relative weights they place on income accruing in the more or less distant future, they also will disagree on which plans maximize the expected present value of profits. This is especially likely to be a problem when the firm is considering investing in a new process or product whose benefits and costs are unknown.

solution

A partial solution to this difficulty is achieved if there are so-called **complete and competitive markets** (a concept investigated in more detail in Chapter 3). Then maximizing the market value of the firm is an appropriate goal to which the owners would agree. In this context, having complete and competitive markets means that any individual can use the financial and insurance markets to move income across time and shift it between different uncertain events, all at given prices. In doing so, the individual achieves whatever patterns of receipts he or she may like and can afford. In such circumstances, it is best to make the firm's value (evaluated at the given market prices) as large as possible, because this provides the largest amount to invest, and this amount can then be invested in light of the person's own preferences and beliefs about the future (see Chapter 14). However, markets are almost surely not complete in this sense. Thus, there are disagreements among owners about the optimal course of action for the firm, and market value maximization does not always win unanimous approval.

Não mercado é competitivo

Agency theory

Third, to the extent that those making the decisions are not the only claimants on returns from the concern, they may wish to maximize their portion of the flows, rather than total returns. For example, suppose the owners are stockholders with limited liability who are at risk for the firm's obligations only up to the amount of their investments, but some of the firm's financing is via debt that must be repaid before the owners get any returns on their investments. In that case, the owners might prefer to make investments that are excessively risky rather than to maximize the value of the firm (including both the value of its debt as well as its equity). When things go well, the debtors are paid their contracted amounts and the equity holders keep the remainder. When things go badly, the debt is not repaid in full and some of the losses are thus shifted onto the debtors. Riskier investments shift more of the downside onto the debtors and leave the upside to the stockholders. Therefore, they might be preferred by the latter group, even if the increased risk reduced the value of the debt by more than it increased the value of the stock. We explore this particular conflict again in Chapter 5 in the context of the savings and loan industry.

## Other Goals and Stakeholders' Interests

Of course, there are many organizations where profit maximization is clearly not the goal. Mutual insurance companies reduce the premiums paid by their policyholders through dividends (distributions of part of the excess of receipts over costs). Customer-owned cooperatives are intended to sell to their members at lower prices than profit maximization would yield. Clearly, assuming profit maximization is likely to be inappropriate for such firms. However, in each of these cases the customers (the policyholders and the co-op members) are the nominal owners, and useful analysis can be conducted using the assumption that the organization is structured to attempt to serve their interests. Similarly, in the case of a firm that is entirely employee owned such as Avis, the automobile rental company, or Egged, the Israeli intercity bus company, an examination of the employees' interests should give insight to the policies that the firm will pursue.

coops - não maximiza lucro e sim atender membros

Even if we are willing to assume that a privately-owned firm will be designed to serve the interests of the owners, however, in whose interests is a research university run? Surely, "in the public interest" is an inadequate answer. The public consists of a multitude of people whose interests may be in conflict. Students may prefer faculty to be selected largely for teaching ability; employers may want teaching that gives students professionally applicable training and research that is geared to quick industrial application; alumni booster clubs might prefer that more resources go to generating winning athletic programs; taxpayers might want subsidies held down and only local students admitted. Generally, not-for-profit organizations have no owners in the usual sense. In such circumstances, predicting organizational form and behavior requires careful analysis of who has the power to design the organization, who can make decisions, and who can influence these decisions and their implementation.

Heterogeneity

A similar difficulty arises in firms in which the interests of **stakeholders** other than the nominal owners are given legitimacy. For example, a survey of the presidents of 100 major Japanese firms asked what the objectives of companies should be, and what their actual goals were.[11] In each case, the pursuit of shareholders' profit came a very distant fourth on the list, garnering only 3.6 percent of the responses. Asked the questions of to whom companies should belong and to whom they actually do belong (with multiple answers allowed), senior executives listed shareholders first among those entitled to ownership, but employees first among the actual owners. A survey of Japanese middle managers indicated that they viewed stockholders as only fourth in a list of those to whom companies should belong, with employees first. Shareholders came third on the middle-level managers' list of the effective owners of the firm, behind employees and management.[12]

Management in other countries also often appears to be concerned with the interests of the nonowner stakeholders, including employees, suppliers, customers, the communities in which the firm is located, and those for whom the natural environment is affected by the firm's actions. To the extent that the nominal owners cannot enforce exclusive attention to their interests in management's decisions, the

[11] Reported in the *Nihon Sangyo Shimbun*, July 5, 1990. We are grateful to Masahiro Okuno-Fujiwara of the University of Tokyo for this and the next reference.

[12] Reported in *Nihon Sangyo Shimbun*, April 23, 1990. Multiple responses were allowed.

actual policies and practices of the firm are going to represent a political compromise mediated by management.

In summary, the assumption that organizations are maximizing entities with well-defined goals is one that should be made very cautiously.

## Modeling Human Motivation and Behavior

A central premise of economic analysis is that people (as opposed to organizations) do have well-defined interests describable by individual utility functions, and that they seek to maximize their utility. Although this assumption is far from uncontroversial, it in fact has virtually no empirical content in light of the limitless factors on which individual utility could depend.

For example, a sufficient concern both for the well-being of others and for social approval could rationalize apparently extreme self-sacrifice. There is no doubt that such factors are both real and, in some cases, immensely important. Although a narrow calculation of personal costs and benefits might usually be enough to prevent soldiers from deserting in the face of battle ("Am I more likely to be shot by the enemy or by a firing squad?"), sacrificing one's life for one's comrades seems hard to explain without appeal to altruism or to an exceedingly high regard for others' opinions of one's courage. Such factors are important even in the more mundane problems that concern us here, such as motivating workers to provide honest effort or providing incentives for borrowers not to skip out on their creditors. Furthermore, important features of many organizations can be best understood in terms of deliberate attempts to change the preferences of individual participants to make these factors more salient. As a result, organizationally desired behavior becomes more likely. This is clearly an element of leadership as it is usually understood, and it has much to do with practices of organizing semipermanent groups of workers and encouraging them to interact socially as well as at work.

### Rationality-Based Theories

While admitting all this and, in particular, the possibility of intrinsic motivation, we adopt the view that many institutions and business practices are designed as if people were entirely motivated by narrow, selfish concerns and were quite clever and largely unprincipled in their pursuit of their goals. This model of human behavior incorporates the sort of rational self-interest usually assumed in standard economic models, such as that of a consumer making utility-maximizing purchases at given prices. But it goes further. It posits that people will be very sharp in discovering even subtle ways in which they can advance their interests and that they will be fundamentally amoral, ignoring rules, breaking agreements, and employing guile, manipulation, and deception if they see personal gain in doing so.

As we shall see, this assumption does have bite: It often serves to give sharp, testable predictions and explanations. Moreover, even though it is an extreme caricature to regard people as amorally motivated solely by narrow self-interest, the predicted institutions and practices are often not very sensitive to this caricature. A bank has guards, vaults, and audits because it would otherwise be robbed; this explanation of practices is unaffected by the observation that many honest people would not rob an unguarded bank. Similarly, organizations design reward schemes so that employees find it in their personal interests to work to advance the organization's goals. The fact that many employees would not instantly abandon the organization's interests even if their incentives were removed is not particularly important for our analysis.

At the same time, we do not automatically make the hyperrationality assumptions common in some economic analyses: that people are capable of instantaneous, unlimited, perfect, and costless calculation, that they can effectively and effortlessly forecast all possible eventualities and the full implications of any information or decision, and that they completely optimize in all situations. These assumptions are not just counterfactual; they also prevent understanding of many important elements of organizations. For example, organizations regularly employ **routines**—standard operating procedures and rules of thumb—for obtaining information, making and implementing decisions, and carrying out tasks. Using routines economizes on scarce and valuable decision-making resources, although it sometimes means that the decisions that are made are not the best ones that could have been reached if the problems were subjected to a full analysis rather than treated routinely. Organizational learning occurs when these routines are modified in response to new knowledge. None of this makes much sense if we assume hyperrationality.

Paradoxically, the very imperfections in the rationality of people and in the adaptability of organizations denied by many simple economic theories are necessary in proving that the rationality-based theories are descriptively and prescriptively useful. With perfect rationality, one would rarely expect to observe two organizations in substantially. the same circumstances making substantially different choices, so there would be no possibility of testing what kinds of organizations perform best. For example, to test whether the commonly observed scheme of paying commissions to compensate insurance agents is an efficient system of sales incentives, we would want to compare these firms to other firms that do not provide such incentives to evaluate their performances, or possibly to the same firm using different practices at different times. It is untenable to adhere too closely to tenets of individual and organizational rationality and at the same time to claim an empirical basis for the theory. A more defensible position, suggested by Richard Nelson and Sidney Winter, is that people learn to make good decisions and that organizations adapt by experimentation and imitation, so that there is at least "fossil evidence" available for testing theories.

Nevertheless, theories based on perfect rationality and adaptability are surprisingly successful in generating explanations and specific predictions about observed institutions and business practices, and they form the main focus of this book.

## Case Study: Coordination, Motivation, and Efficiency in the Market for Medical Interns

The evolution of the system under which graduating medical students in the United States are matched with hospitals seeking interns and residents provides a striking illustration of a number of the themes and concepts introduced in this chapter.[13]

### Matching Problems and Failed Solutions

The practice of new M.D. degree holders taking internships in hospitals as a clinical stage in their medical education appeared in the United States about the beginning of the twentieth century. It gave the interns practical training and the hospitals cheap labor. Until graduates of foreign medical schools began seeking U.S. internships in significant numbers in the 1970s, the number of positions open in hospitals for interns always exceeded the number of students seeking internships, so competition for interns

[13] This section is based on two papers by Alvin E. Roth, "The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory," *Journal of Political Economy*, 92 (December 1984), 991–1016, and "A Natural Experiment in the Organization of Entry Level Labor Markets: Regional Markets for New Physicians in the U.K.," *American Economic Review*, 81 (June 1991), 415–40. See also Roth's less technical discussion, "New Physicians: A Natural Experiment in Market Organization," *Science*, 250 (December 14, 1990), 1524–28.

was intense. Of course, students differ in their overall attractiveness and in their specific appeal to different schools, so the competition was more intense for some than for others. Moreover, hospitals seeking several interns might be concerned with the particular mix of students they got, and not just with the overall quality of the individuals they attracted. They would consequently have preferences over groups of students. At the same time, students may have differing individual preferences among the hospitals at which they might intern. In sum, the final matching of students and hospitals is of great concern to all.

Various methods were used to match students and hospitals over the years. First was a scheme similar to that used in U.S. college and graduate school admissions today. Students would apply to hospitals that were seeking interns. The hospitals then ranked their applicants and offered positions to some number while telling others that they were alternates on a wait-list. The students then had to decide whether to take one of their firm offers or to wait to see if a hospital they preferred would make them a firm second-round offer. The problem was that the hospitals began competing by making their offers earlier and earlier, sometimes as early as only half-way through the students' medical school studies. This was very bad from the point of view of the hospitals, because at that stage they could learn so little about the students' ultimate performance and interests. However, each hospital still had an individual incentive to accelerate its offers to try to recruit the best students. The process was also very disruptive for the medical schools and the students' studies.

The eventual response was for the medical schools to agree not to give out information on their students until late in their careers. This moved the recruiting into the students' final year of medical school, but a new problem arose. Students who had been accepted by one of their less-preferred hospitals but wait-listed by one they liked better would hold off responding to the first hospital's offer as long as possible, hoping that they would clear the other hospital's wait-list. Again, this was individually optimal behavior by all concerned, but it meant that there were student-hospital matches that were missed but that were mutually preferred to those that actually occurred. This in turn led to students' reneging on their acceptances.

Attempts were made to overcome these problems. The chief one was for the hospitals to agree on the earliest date for making offers, for them to shorten the time given to students to make their decisions (down to as little as a twelve-hour period from midnight to noon!) and to limit communication between the hospitals and students during the period that offers were outstanding. Perhaps unsurprisingly, these measures proved futile, and the turmoil continued.

### The National Intern Matching Program

In 1951, a centralized scheme was introduced by the hospitals and medical schools on an experimental basis to coordinate the matching of students to internships. The workings of the experimental version were adjusted in light of students' observations that they would do better under its rules to misrepresent their preferences among hospitals, and the revised scheme was formally instituted in 1952. Under it, students seeking internships and hospitals seeking interns first exchange information with one another, much as before. Then the students rank the hospitals in which they are willing to be employed and the hospitals similarly rank the groups of student applicants whom they are willing to take. These rankings are then submitted to a central office, which uses a specific rule (algorithm) to match students and hospitals. This new system, called originally the National Intern Matching Program (NIMP), ended the turmoil that had marked the market for interns. It has remained in place, drawing the voluntary participation of the vast bulk of hospitals and medical students (up to 95 percent of students). However, beginning in the mid-1970s there was an increasing number of married couples needing two internships in the same vicinity, and large numbers of these people sought and found matches outside the NIMP system.

The actual algorithm used to match students and hospitals is somewhat complicated, and understanding its detailed workings is not necessary for what follows. However, the basic idea can be explained simply for the (unrealistic) case in which each hospital has space for only one intern and there is an equal number of students and hospitals.[14]

The algorithm works through rounds in which it seeks to match students and hospitals on the basis of their submitted preference orders. In essence, each hospital at each round offers its position to its most preferred applicant. Students offered a place are then tentatively matched with their favorite among the hospitals making them an offer. Their names are then stricken from the submitted rankings of all those hospitals which they ranked lower than the one with which they are tentatively matched, and the process is repeated using the hospitals' revised rankings. Note that generally some students will have moved up in some of the hospitals' revised rankings, because other students whom these hospitals actually liked better have been (tentatively) matched elsewhere. This means that students who have moved to the top of some hospital's rankings will now get offers from that hospital. If they were already tentatively matched with some other hospital but prefer the new offer, the old tentative match is broken and they are (still tentatively) assigned to their preferred choice. The process then continues until everyone is tentatively assigned, at which point the assignments are made final and announced.[15]

EFFICIENCY AND STABILITY Consider now the general case where the hospitals have multiple slots for interns and there is no presumed equality between the numbers of students and available spaces. Even in this context, the matches that the NIMP generates can be shown to be efficient in the usual sense: There is no way to reassign the students to the hospitals so as to make one of the students or one of the hospitals better off without hurting some other hospital or student. This was not necessarily true of the older systems it replaced. Moreover, the NIMP avoids many of the transaction costs that marked the older systems. The process is relatively straightforward, and most students and hospitals have accepted its results rather than trying to go around it to find mutually preferred matches.

However, in the present context, monetary transfers and "side payments" are not permitted. Thus, efficiency cannot be equated with value maximization because compensation cannot be paid. There may then be many different efficient patterns of matches. In light of this, it is especially significant that the matching generated by the NIMP can be shown to enjoy an extra property that is much stronger than efficiency in this context: the NIMP matches are *stable*. Specifically, there will never be a hospital-student pair where the student prefers the given hospital to the one to which the algorithm assigned him or her and the hospital would rather have the given student than one of the students actually assigned to it. Thus, recontracting between individual students and hospitals cannot upset the outcome of the process, even when the hospitals are free to drop people who have been assigned to them and students are free to ignore the assignment they have been given.

Stability means that the proposed assignment is not only efficient for the group as a whole, it is also efficient for every subgroup, even if that group ignores the effects

[14] We are indebted to Robert B. Wilson of Stanford University for this interpretation.

[15] Note that each student gets a more preferred hospital whenever his or her tentative match is changed. Thus, with only a finite number of hospitals, the algorithm eventually must cease generating changes, at which point it stops.

Table 2.1 An Example of Student and Hospital Preferences

| Ranking | Student |  |  | Hospital |  |  |
|---|---|---|---|---|---|---|
|  | Alice (A) | Barbara (B) | Charlie (C) | Hopkins (H) | Stanford (S) | Yale (Y) |
| 1st | Y | S | S | A | A | B |
| 2nd | S | Y | Y | B | B | A |
| 3rd | H | H | H | C | C | C |

of their decisions on non-members. Regardless of what group of students and hospitals form to seek alternative matches, the members can never find a set of matches that involves only members of the group that is better for all of them than the match initially proposed by the NIMP. Stability is a very demanding condition. The fact that the NIMP matches are stable helps to explain the persistence of the program.

It may seem remarkable that the hospitals and medical schools were able to devise an efficient, stable system. What is perhaps more remarkable is that this scheme is ideal from the hospitals' point of view (and quite the opposite from the students' perspective).

If a hospital wants some specific number *n* of interns, then the NIMP algorithm actually assigns that hospital the *n* students it ranks highest among all the students it could ever get at one or another stable assignment (no matter how computed)! So if one stable assignment gives a hospital its first and fourth choices and another gives it its second-ranked and third-ranked students, the NIMP actually gives it its first- and second-ranked students. At the same time, each student is assigned to the hospital he or she ranks *lowest* among all those to which he or she could be assigned at *some* stable assignment.

All this presumes that the students and hospitals submit lists reflecting their actual, true rankings. In this regard, the NIMP consistently claims that there is no advantage to misrepresenting preferences. In fact, no student or hospital can gain by misrepresenting its first choice. However, it is not individually optimal for students to list their complete true preferences independent of the hospitals' rankings and the lists being submitted by other students, or for a hospital with more than one slot to fill to submit an honest list independent of the students' lists and those of the other hospitals.

An Example of Successful Strategic Misrepresentation To see how misrepresentation could theoretically be beneficial, consider an example that is remarkably simple compared to the real situation faced by actual students.[16]

Suppose there are three students, Alice, Barbara, and Charlie, and three hospitals, Hopkins, Stanford, and Yale. Each hospital wants only one student, so by the results claimed earlier, they have no reason to misrepresent their preferences. The assumed rankings are given in Table 2.1.

At the first round, Hopkins and Stanford offer positions to Alice, the most preferred student, and Yale offers one to its favorite, Barbara. Alice is tentatively matched with Stanford, which she prefers to Hopkins, and Barbara is tentatively matched with Yale. Thus, both Alice and Barbara are removed from Hopkins's rankings, moving Charlie to the top (by default as he is the only one left). At the next round, Stanford and Yale repeat their offers to Alice and Barbara, respectively, because each is at the top of the corresponding school's list, and Hopkins makes an offer to Charlie. All the students and hospitals are now matched, and these matches are the ones actually announced. Alice and Barbara end up with their second-favorite hospitals, but these two hospitals get their first picks.

Table 2.2 Hospital's Revised Rankings at Round III of the Algorithm

| Ranking | Hospital |  |  |
|---|---|---|---|
|  | H | S | Y |
| 1st | B | A | A |
| 2nd | C | B | C |
| 3rd | — | C | — |

Now suppose that Barbara misrepresents her preferences and submits a list that still ranks Stanford first, but then puts Hopkins rather than Yale in second place. Tracing through the workings of the computer algorithm is much more complicated in this case (and you may prefer to skip the next two paragraphs, which describe the process), but it turns out that this misrepresentation will work to Barbara's advantage. She ends up being assigned to her first choice, Stanford.

At the first round, Alice is again tentatively matched with Stanford and Barbara with Yale. Now, however, since Barbara claims that she views Yale as third best, her name is not removed from Hopkins's list. Alice's name is, however, still taken off Hopkins's list, which now lists Barbara first and Charlie second. At the next round, Hopkins and Yale make offers to their now-favorite, Barbara, and Stanford repeats its offer to Alice. Barbara claims to like Hopkins better than Yale, so her tentative match with Yale is broken, and she is matched provisionally with Hopkins. As a result, she is removed from Yale's ranking. The revised rankings of the hospitals going into the third round are given in Table 2.2.

At the third stage, each hospital again makes an offer to the top-ranked student on its revised listing: Hopkins to Barbara, and Stanford and Yale to Alice. The tentative matches set by the computer are Barbara again with Hopkins, but Alice with her top choice, Yale. Thus, Alice is dropped from Stanford's list, moving Barbara to the top. Now, at the fourth stage, Barbara gets offers from both Hopkins and Stanford, while Alice again is Yale's choice. The computer now matches Barbara with Stanford, her first choice, and removes her from Hopkins's list. Consequently, at the fifth round, the computer matches Alice with Yale, Barbara with Stanford, and Charlie with Hopkins. This is the match that is announced.

Barbara's misrepresentation allowed her to get her first choice rather than her second, as she would have gotten by being honest. (Incidentally, it also moved Alice from her second-ranked hospital to her first.) By downgrading the hospital that ranked her first and would offer her first-round admission, Barbara manages to stay on Hopkins's list after being first matched with Yale, while Alice is dropped from Hopkins's list. Then Barbara is offered admission by Hopkins, which allows her to eliminate herself from Yale's list. This leads Yale to take Alice, opening the spot at Stanford for Barbara. Figuring out an advantageous preference misrepresentation was not obviously straightforward, since it depended on knowing the rankings of both the hospitals and the other students as well as the workings of the algorithm, but one existed.

In fact, medical students do spend time and energy trying to learn what the

[16] We are indebted to Jeremy Bulow of Stanford University for this example.

various hospitals are looking for, what their historical patterns of rankings have been, and what other students are likely to do, hoping thereby to figure out an advantageous way to misrepresent their preferences. However, as the example suggests, figuring out a worthwhile strategic misrepresentation is subtle and difficult. It is quite possible that students eventually give up trying to game the system and report accurately.

### The Evolution and Persistence of Organizational Forms

The logic introduced in this chapter suggests that the efficiency and stability of the NIMP have contributed crucially to its survival, just as the failure of the earlier systems can be attributed to their inefficiency and instability. When an inefficient arrangement is in place, there is a general interest in supplanting it with one that will make everyone better off. When the arrangements are unstable in the sense we have discussed here, they are particularly fragile because pairs of agents have both the incentive and the ability to subvert their workings and presumably will do so.

This argument is supported by consideration of the mechanisms that have been used in the United Kingdom to address the problem of matching new physicians and hospitals there. The situation in the United Kingdom is more complex than in the United States, because there are seven regional markets, each of which has tried its own matching algorithms after a period of instability similar to that experienced in the United States before 1951. Eight of these different algorithms have been analyzed formally, of which two were found to be stable and six were not. The two stable ones are still in use, while four of the six unstable ones have been discarded.

As noted earlier, an increasing number of student couples in the United States are finding internships for themselves outside the NIMP. In terms of the theory proposed here, a reason for this is that the NIMP algorithm is not guaranteed to produce stable matches when couples look for assignments together. Thus, the recontracting problems that beset the older systems are effectively reappearing. In fact, no algorithm is guaranteed to find stable assignments in these environments, because none need exist! This suggests that instability in this market may become endemic.

## Summary

There are economic organizations at many levels, from the economy as a whole to firms to units within them. Within the theory, the firm is distinguished from other, smaller units by its status as a legal entity able to enter into binding agreements with individuals. This power makes it unnecessary for the individuals to enter into a complex multilateral contract to organize their transactions and consequently makes it more likely that efficient arrangements can be negotiated.

The basic unit of analysis in economic organization theory is the transaction, where goods or services are transferred from one person to another. An important focus of the analysis is on the behavior of the individuals who transact. The main tasks of economic organization are to *coordinate* the actions of the various individual actors so that they form a coherent plan and to *motivate* the actors to act in accordance with the plan.

We evaluate organizations on the basis of how well they satisfy the wants and needs of people, that is, on the basis of their *efficiency*. Since organizations are partially designed, one can also explain features of organizations as attempts by the organization designers to achieve efficiency. The relative successes of different kinds of organization provide some of the main evidence for theories about which kinds of organization are most efficient in particular environments.

*Using efficiency as a positive principle* requires taking care about whose interests are being served and what kinds of arrangements are feasible. A small group that is able to bargain among its members may decide on arrangements that are efficient for themselves but that would be regarded as inefficient if the group could be enlarged. Efficiency in this sense is used to make predictions only and not to evaluate the social desirability of the agreed arrangements. In addition, arrangements that appear to be wasteful may still be efficient in the positive sense if there is no feasible alternative for the group that all would prefer.

The expansion of production in modern economies has been accomplished in large measure through *specialization*, according to which any one individual performs only a tiny fraction of the kinds of tasks required to make what he or she uses. Increased specialization implies that people become more reliant on the work of others, and the need for coordination increases. The two extreme alternative ways to coordinate are to communicate information to a central planner who makes all the important decisions or to provide individuals with the information and resources they need to make decisions that fit in with the overall plan. Both extremes are mere caricatures. Real economies all use a mix of these two approaches.

Transaction costs are the costs of negotiating and carrying out transactions. They include *coordination costs*, such as the costs of monitoring the environment, planning and bargaining to decide what needs to be done, and *motivation costs*, such as the costs of measuring performance, providing incentives, and enforcing agreements to ensure that people follow instructions, honor commitments, and keep agreements.

The way the transactions are best organized and managed depends on the basic attributes of the transaction. Five attributes have been identified as especially important. The first is *asset specificity*. When parties are called upon to make larger specific investments, they generally seek to organize in ways that safeguard those investments. Second, when one party is involved in *frequent, similar transactions* over a *long duration*, it is likely to pay that party to set up more specialized mechanisms or procedures to reduce the costs of transacting. Frequent transacting between two or more people over a long horizon allows the parties to develop understandings, reducing the need for explicit agreements, and to grant or withhold favors, reducing the need for outside enforcement of agreements. Third, *uncertainty* about the circumstances in which a transaction will occur and the *complexity* of the decisions that will be required make it hard to forecast exactly what performance will be required. This undermines the effectiveness of simple contracts and leads parties to contract over decision rights and procedures rather than over specific aspects of performance. Fourth, the *costliness of measuring performance* makes it difficult to provide performance incentives, leading the parties to seek organizations where measurement and incentive issues are of less importance. Fifth, when a transaction is *closely connected to other transactions*, that is, where failures of close fit among the transactions are quite costly compared to failures to make best use of local resources, coordination mechanisms tend to be strengthened, either by increasing managerial oversight or by arranging frequent meetings among the people responsible for the individual transactions.

In its simplest form, transaction costs theory holds that organizations are designed to minimize the total costs of transacting. The two problems with this simple theory are that the costs of transacting are not logically distinguishable from other costs, and that efficiency itself does not always imply total cost minimization. However, there is one special case where the latter problem, at least, disappears.

When individual preferences are free of wealth effects, that is, when everybody regards each outcome as being completely equivalent to receiving or paying some amount of money and when there are no a priori restrictions on monetary transfers, the efficient allocations are precisely those that maximize the total value and divide it all among the participants. This conclusion is known as the *value maximization principle*. The *Coase theorem* holds that when there are no wealth effects, all decisions about productive activities and organizational arrangements are unaffected by the wealth, assets, or bargaining power of the parties. Only the decision of how benefits and costs are to be shared is affected by these factors. This view contrasts, for example, with the Marxian view that organizations reflect underlying power arrangements and class interests and not the desire to maximize total wealth. When there are no wealth effects, an efficient organization acts as if it were an individual with a well-defined objective to maximize total value.

In the general case, the value maximization criterion does not describe how organizations behave. Organizations then may serve a variety of conflicting individual interests, rather than maximizing a single overall organizational objective. This is especially true of public organizations, like universities, with their ever-shifting balance among different social interests, but it is also true in varying degrees about business firms, where even the owners may have divergent interests.

While we do not attribute motives to organizations in general, we do attribute them to people. In the theories treated in this book, people are self-interested and opportunistic, and successful organizations must channel that self-interest into socially beneficial behavior.

## ■ Bibliographic Notes

As with so many of the central concerns of economics, the problems of economic organization and organizations find their first treatment of lasting significance in Adam Smith's *The Wealth of Nations*. Although organizational issues were not a major focus of mainstream economists after Smith, important insights come from the writings of Karl Marx in the nineteenth century and especially from Frank Knight and John Commons in the first quarter of the twentieth century. Commons in particular championed the treatment of the transaction as the fundamental unit for analysis, while Knight specifically addressed the organization of firms, and of economic activity more generally, in efficiency terms.

Ronald Coase is rightly viewed as the originator of transaction cost economics, and much of what we have reported in this chapter is an outgrowth of his classic 1937 paper in which he first developed the idea that economizing on transaction costs would determine the organization of economic activity and the division of activity between firms and markets. His 1960 paper, which develops the Coase Theorem, is another classic. It has done much to make economists aware of the power of using value maximization and efficiency as positive, explanatory principles. These papers were specifically cited when Coase was awarded the Nobel prize in Economics in 1991.

The importance of dispersed, local information for economic organization was accentuated by Friedrich Hayek in his contribution to the debate about market systems versus central planning that followed the establishment of the centralized communist system in the USSR.

Among more recent contributions, the contractual approach to organizations was championed by Armen Alchian and Harold Demsetz in seeking to explain the role of hierarchy and supervision in the firm in incentive terms. Kenneth Arrow's influential little book develops his treatment of organizations as arising when markets fail. Oliver Williamson's writings have played a major role in developing transaction cost economics. His 1985 book gives an excellent overview of his approach, which identifies asset specificity, frequency, and uncertainty as the key dimensions of transactions and which also accentuates the limits of human rationality. This latter theme was first introduced into economics by Herbert Simon, and has been developed in an evolutionary direction by Richard Nelson and Sidney Winter. Yoram Barzel, building on the contributions of Stephen Cheung, has emphasized the measurement costs dimension of transaction cost economics. The connectedness dimension and the notion of design connectedness are introduced for the first time here.

The survey papers by Bengt Holmstrom and Jean Tirole and by Williamson are valuable supplements not only for the questions in this chapter, but for many of the issues addressed throughout the book. Our paper on bargaining and influence costs is an integrated presentation and critique of the basics of transaction cost economics.

## ■ References

Alchian, A., and H. Demsetz. "Production, Information Costs, and Economic Organization," *American Economic Review*, 62 (1972), 777–95.

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

Barzel, Y. "Measurement Costs and the Organization of Markets," *Journal of Law and Economics*, 25 (1982), 27–48.

Cheung, S.N.S. "Transaction Costs, Risk Aversion, and the Choice of Contractual Arrangements," *Journal of Law and Economics*, 12 (1969), 23–42.

Coase, R. "The Nature of the Firm," *Economica*, 4 (1937), 386–405.

Coase, R. "The Problem of Social Cost," *Journal of Law and Economics*, 3 (1960), 1–44.

Commons, J.R. "Institutional Economics," *American Economic Review*, 21 (1931), 648–57.

Hayek, F.A. "The Use of Knowledge in Society," *American Economic Review*, 35 (1945), 519–30.

Holmstrom, B.R., and J. Tirole. "The Theory of the Firm," Chapter 2 in R. Schmalensee and R. Willig, eds., *Handbook of Industrial Economics* (New York: North-Holland, 1989).

Knight, F.H. *Risk, Uncertainty and Profit* (London: London School of Economics, 1921).

Marx, K., *Capital* (Harmondworth, UK: Penguin Books, 1976).

Milgrom, P., and J. Roberts. "Bargaining and Influence Costs and the Organization of Economic Activity," in J. Alt and K. Shepsle, eds., *Perspectives on Positive Political Economy* (Cambridge: Cambridge University Press, 1990).

Nelson, R.R. and S. Winter. *An Evolutionary Theory of Economic Change* (Cambridge, MA: Harvard University Press, 1982).

Simon H. *Models of Man* (New York: John Wiley & Sons, 1957).

Smith, A. *An Inquiry into the Nature and Causes of the Wealth of Nations* (Oxford: The Clarendon Press, 1976).

Williamson, O. *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting* (New York: The Free Press, 1985).

Williamson, O. "Transaction Cost Economics," Chapter 3 in R. Schmalensee and R. Willig, eds., *Handbook of Industrial Economics* (New York: North-Holland, 1989).

## EXERCISES

### Food for Thought

1. One of the main tenets of economic analysis is that people act in their own narrow interests. Why, then, do people leave tips in restaurants? If a study were to compare the size of the tips earned by servers in roadside restaurants with those frequented mostly by locals, what would you expect to find? Why?

2. For most large Japanese firms, a majority of the voting shares are owned by the company's major lenders (banks and insurance companies), its customers and suppliers, and related firms with whom it has longstanding relationships. How would this ownership structure affect the objectives that these firms seem to pursue?

3. Would you expect the organization of agriculture in developing countries to be arranged in a way that maximizes the total wealth of the farmers, workers, and lenders? If arrangements do not maximize total wealth, what kind of variations would be most likely? Explain your answer.

4. In California's fruit farms, farm workers who pick fruit are commonly organized into teams that are paid according to the number of trees that are cleanly picked. The teams themselves decide how to divide the pay among their members. What attributes of this transaction account for this arrangement?

5. Cable television companies lay cables to individual households in the communities they serve to carry the television signal. How specific is this investment? What kind of arrangements would you expect the cable companies to make with local communities about the pricing and taxation of cable services?

### Mathematical Exercises

1. Suppose four families share a common stretch of beach and they are considering a program of improvements, including a stairway and a play structure for children. The value of spending $y$ in total for the improvements is $5y-\frac{1}{2}y^2$ for families #1 and #2, $7y-\frac{1}{2}y^2$ for family #3, and $4y-y^2$ for family #4. What is the efficient level of expenditure on beach improvements?

2. Continuing with the situation hypothesized in problem 1, show that if the cost of the improvements is to be shared equally, then family #4 will be unwilling to bear its share of the cost. What is the largest improvement that all the families would agree to if the cost of improvement must be shared equally? Demonstrate that the resulting expenditure is inefficiently low by finding an alternative level of expenditure and pattern of cost sharing that the families unanimously prefer.

3. (Mathematical proof of the value maximization principle.) Suppose that there are $N$ individuals. Person $n$'s utility when outcome $y$ occurs and he or she receives cash compensation of $x_n$ is given by the utility function $x_n + v_n(y)$. Suppose that decision $y$ generates net profits $P(y)$, which may be a positive number if we think of $y$ as representing an investment or a negative number if we think of $y$ as representing some public good, such as parks or roads, at a total cost of $-P(y)$. The profits are divided among the individuals with individual $n$ receiving $x_n$ (or paying $-x_n$). The payments must add up to the amount available, that is, $P(y) = x_1 + \ldots + x_N$. Prove that an allocation $(y,x_1,\ldots,x_N)$ is efficient if and only if $y$ maximizes the total value $P(y) + v_1(y) + \ldots + v_N(y)$. [*Hint:* There are two things to prove. First, you must show that if an allocation does maximize the total value, then there cannot be an allocation that Pareto dominates it. Second, if the allocation does not maximize the sum, then you must show that there is another allocation that dominates it. To show the latter, take any $y$ with a higher total value and show that the $x_n$s can be chosen so that the gain in total wealth is divided equally among the participants.]

4. (Characterizing a utility function when there are no wealth effects.) Suppose that a decision maker's preferences are such that for any two decisions $y$ and $y'$, there is an amount of cash compensation or a cash payment such that $y$ combined with cash compensation of $C(y,y')$ would be just as good, from the decision maker's perspective, as $y'$ with a zero-cash compensation. Further suppose that this amount $C(y,y')$ does not depend on the level of other cash payments made to or by the decision maker. Finally, suppose that the decision maker prefers more money to less. Fix any possible decision $\bar{y}$ and define $v(y) = C(\bar{y},y)$. Show that with this definition, the utility function $x + v(y)$ represents the decision maker's preferences; that is, the decision maker will prefer an allocation $(x,y)$ to another allocation $(x',y')$ if and only if $x + v(y) > x' + v(y')$. [*Hint:* Argue first that the decision maker is indifferent about having $(x,y)$ or $(x+v(y),\bar{y})$. Therefore, for any utility function that represents the decision maker's preferences, $U(x,y) = U(x+v(y),\bar{y})$ and, similarly, $U(x',y') = U(x'+v(y'),\bar{y})$.]