For Discussion

1. Assume that the world market for crude oil is competitive, with an upward-sloping supply schedule and a downward-sloping demand schedule. Draw a diagram that shows the equilibrium price and quantity. Now imagine that one of the major oil exporting countries undergoes a revolution that shuts down its oil fields. Draw a new supply schedule and show the loss in consumer surplus in the world oil market resulting from the loss of supply. What assumptions are you making about the demand for crude oil in your measurement of consumer surplus?
2. Now assume that the United States is a net importer of crude oil. Show the impact of the price increase resulting from the loss of supply to the world market on social surplus in the U.S. market.

5

Rationales for Public Policy

Market Failures

The idealized competitive model produces a Pareto-efficient allocation of goods. That is, the utility-maximizing behavior of persons and the profit-maximizing behavior of firms will, through the "invisible hand," distribute goods in such a way that no one could be better off without making anyone else worse off. Pareto efficiency thus arises through voluntary actions without any need for public policy. Economic reality, however, rarely corresponds perfectly to the assumptions of the idealized competitive model. In the following sections we discuss violations of the assumptions that underlie the competitive model. These violations constitute market failures, that is, situations in which decentralized behavior does not lead to Pareto efficiency. Traditional market failures are shown as circumstances in which social surplus is larger under some alternative allocation to that resulting under the market equilibrium. Public goods, externalities, natural monopolies, and information asymmetries are the four commonly recognized market failures. They provide the traditional economic rationales for public participation in private affairs.

Public Goods

The term *public*, or *collective*, *good* appears frequently in the literature of policy analysis and economics. The blanket use of the term, however, obscures important differences among the variety of public goods in terms of the nature of the market failure and, consequently, the appropriate public policy response. We begin with a basic question that should be raised when considering any market failure: Why doesn't the market allocate this particular good efficiently? The simplest approach to providing an answer involves contrasting public goods with private goods.

Two primary characteristics define private goods: rivalry in consumption and excludability in ownership and use. *Rivalrous consumption* means that what one consumes cannot be consumed by another; a perfectly private good is characterized by complete rivalry in consumption. *Excludable ownership* means that one has control over use of the good; a perfectly private good is characterized by complete excludability. For example, shoes are private goods because when one wears them no one else can (rivalrous consumption) and, because when one owns them, one can determine who gets to wear them at any particular time (excludable ownership).

Public goods, on the other hand, are, in varying degrees, *nonrivalrous* in consumption, *nonexcludable* in use, or *both*. In other words, we consider any good that is not purely private to be a public good. A good is nonrivalrous in consumption when more than one person can derive consumption benefits from some level of supply at the same time. For example, a particular level of national defense is nonrivalrous in consumption because all citizens benefit from it without reducing the benefits of others—a new citizen enjoys benefits without reducing the benefits of those already being defended. (Each person, however, may value the uniformly provided level of defense differently.) A good is nonexcludable if it is impractical for one person to maintain exclusive control over its use. For example, species of fish that range widely in the ocean are usually nonexcludable in use because they move freely among regions such that no individual can effectively exclude others from harvesting them.

In practice, a third characteristic related to demand, the potential for *congestion*, or *congestibility*, is useful in describing public goods. (The term *crowding* is often used interchangeably with congestion.) Levels of demand for a good often determine the extent to which markets supply public goods at inefficient levels. *A good exhibits congestion if the marginal social cost of consumption exceeds the marginal private cost of consumption*. For example, a considerable number of people may be able to hike in a wilderness area without interfering with each other's enjoyment of the experience (a case of low demand) so that the marginal social cost of consumption equals the marginal private cost of consumption and there is no congestion. However, a very large number of hikers may reduce each other's enjoyment of the wilderness experience (a case of high demand) such that the marginal social cost of consumption exceeds the marginal private cost of consumption and, therefore, the wilderness is congested. The key point to recognize is that some goods may only be nonrivalrous over some range of usage, but at some higher level of use, consumers begin to impose costs on each other. Later we interpret the divergence between social and private marginal costs as an externality.

Excludability and Property Rights

Excludability implies that some individual can exclude others from use of the good. In most public policy contexts in developed democracies, power to exclude others from use of a good is dependent on property rights granted and enforced by the state and its (judicial) organs. *Property rights* are relationships among people concerning the use of things.[1] These relationships involve claims by rights holders that impose duties on others. In contrast to political systems in which institutional arrangements create and enforce property rights, there are anarchic, or Hobbesian, situations with no government and no constraining norms or conventions where physical force alone determines the power to exclude.

Property rights to a good can be partitioned in multiple ways beyond what we think of as simple ownership.[2] For example, a farmer may have a property right that allows using water from a river only during specific months of the year. However, for the purposes of a discussion of excludability, we treat goods as being controlled or "owned" by a single actor. Effective property rights are characterized by clear and complete allocation of claims and high levels of compliance by those who owe the corresponding duties. In this context, where the claim is to exclusive use of the good, compliance simply means accepting exclusion. *De jure* property rights, which are granted by the state, are typically clear though sometimes incomplete. These de jure property rights, however, may be attenuated, or in some cases superseded, by extra-legal behaviors such as trespass, squatting, poaching, or custom. Behaviors such as these may give rise to *de facto* property rights, the claims that actually receive compliance from duty bearers. Sometimes de jure property rights do not exist because changes in technology or relative prices create new goods that fall outside of existing allocations. For example, advances in medical technology have made stem cells valuable, but the right to their use is yet unclear.[3] Note that de facto property rights may or may not present excludability problems, though often they involve substantial costs to the claimants if they must employ physical protection systems, vigilance, stealth, or retaliation to enforce them. If these costs become too high, then individuals may abandon attempts to exclude others from use of the good. In these cases, the good is effectively nonexcludable.

Nonrivalrous Goods

As we concluded in our discussion of efficient pricing, the production of a private good, involving rivalrous consumption, will be in equilibrium at a level where price equals marginal cost ($P = MC$). On the demand side, the marginal benefits consumers receive from additional consumption must equal price ($MB = P$) in equilibrium. Therefore, marginal benefits equal marginal cost ($MB = MC$). Essentially the same principle applies to nonrivalrous goods. Because all consumers receive marginal benefits from the additional unit of the nonrivalrous good, however, it should be

[1]For a seminal discussion of property rights, see Eirik Furubotn and Svetozar Pejovich, "Property Rights and Economic Theory: A Survey of Recent Literature," *Journal of Economic Literature* 10(4) 1972, 1137–62. For a review of property right issues, see David L. Weimer, ed., *The Political Economy of Property Rights: Institutional Change and Credibility in the Reform of Centrally Planned Economies* (New York: Cambridge University Press, 1997), 1–19.

[2]Yoram Barzel, *Economic Analysis of Property Rights* (New York: Cambridge University Press, 1989).

[3]See, Patti Waldmeir, "The Next Frontier: The Courtroom," *Scientific American* 293(1) 2005, 17.

produced if the sum of all individual consumers' marginal benefits exceeds the marginal cost of producing it. Only when output is increased to the point where the sum of the marginal benefits equals the marginal cost of production is the quantity produced efficient.

The sum of marginal benefits for any level of output of a purely nonrivalrous public good (whether excludable or not) is obtained by vertically summing the marginal benefit schedules (demand schedules) of all individuals at that output level.[4] Repeating the process at each output level yields the entire social, or aggregate, marginal benefit schedule. This process contrasts with the derivation of the social marginal benefit schedule for a private good: individual marginal benefit schedules (demand schedules) are added horizontally because each unit produced can only be consumed by one person.

Figure 5.1 illustrates the different approaches. Panel (a) represents the demand for a rivalrous good; panel (b) the demand for a nonrivalrous good. In both cases the demand schedules of the individual consumers, Jack and Jill, for the good appear as D_1 and D_2, respectively. The market demand for the rivalrous good, the downward-sloping dark line, results from the horizontal addition of individual demands D_1 and D_2 at each price. For example, at price P_0, Jack demands Q_1 and Jill demands Q_2, so that the total quantity demanded is Q_0 (equal to $Q_1 + Q_2$). Repeating this horizontal addition at each price yields the entire market demand schedule. Note that at higher prices, above P_{c1}, Jack's *choke price*, only Jill is willing to purchase units of the good, so that the market demand schedule follows her demand schedule. If there were more consumers in this market, then the quantities they demanded at each price would also be included to obtain the market demand schedule.

Panel (b) presents the parallel situation for the nonrivalrous good, with the caveat that we are unlikely to be able to observe individual demand schedules for public goods (for reasons we discuss below). Here, the social marginal benefit at Q_0 (MB_{0S} corresponding to the point on D_S above Q_0) is the sum of the marginal benefits enjoyed by Jack (MB_{01} corresponding to the point on D_1 above Q_0) and Jill (MB_2 corresponding to the point on D_2 above Q_0). The social marginal valuation at this point is obtained by summing the amounts that Jack and Jill would be willing to pay for the marginal unit at Q_0. The entire social marginal valuation schedule (MB_S) is obtained by summing Jack and Jill's marginal valuations at each quantity. Note that for quantities larger than Q_{c1} Jack places no value on additional units so that the social demand (social marginal benefit) schedule corresponds to the demand schedule of Jill.

Notice that in this example the upward-sloping supply schedule indicates that higher output levels have higher marginal costs (a brighter streetlight costs more to operate than a dimmer one); it is only the marginal cost of consumption that is equal, and zero, for each person (they can each look at what is illuminated by the light without interfering with the other). In other words, there are zero marginal social costs of consumption but positive marginal costs of production at each level of supply.

A crucial distinction between rivalrous and nonrivalrous goods is that the valuations of individual consumers cannot directly tell us how much of the nonrivalrous good should be provided—only the sum of the valuations can tell us that. Once an output level has been chosen, every person must consume it. Therefore, the various values different persons place on the chosen output level are not revealed by their purchases as they would be

[4]Paul A. Samuelson, "Diagrammatic Exposition of a Theory of Public Expenditure," *Review of Economics and Statistics* 37(4) 1955, 350–56.

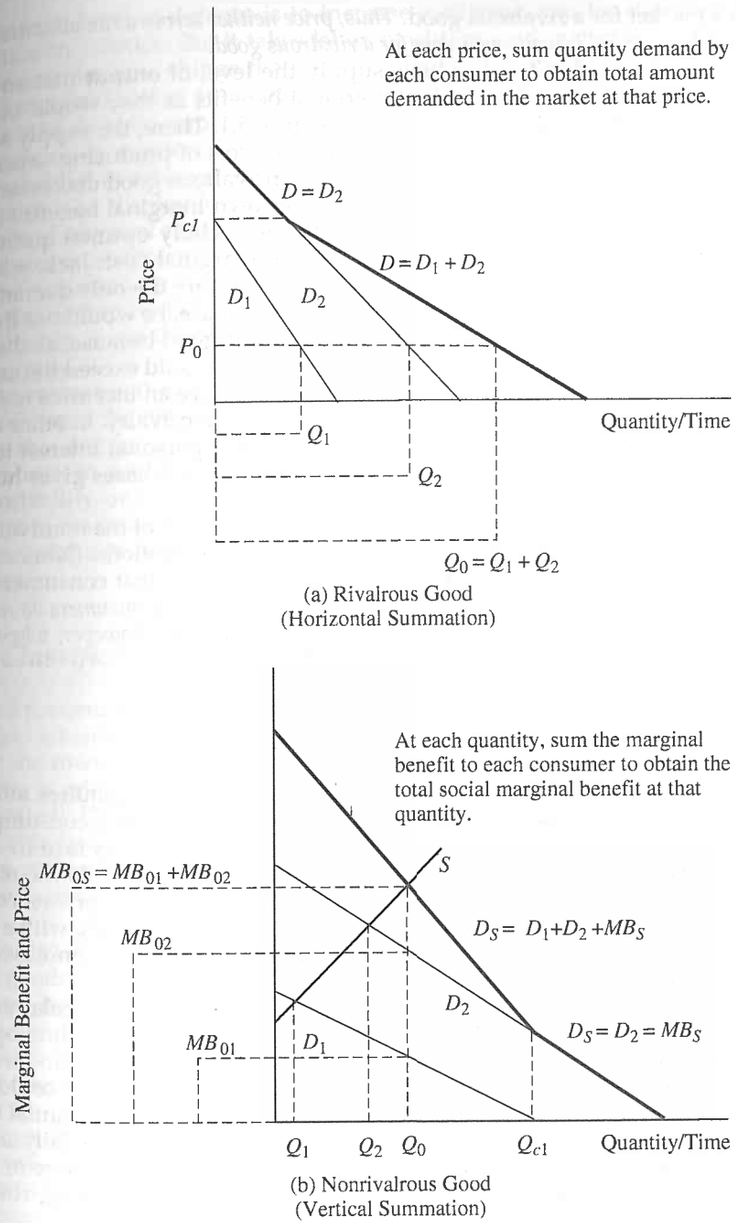


Figure 5.1 Demand Summation for Rivalrous and Nonrivalrous Goods

in a market for a rivalrous good. *Thus, price neither serves as an allocative mechanism nor reveals marginal benefits as it does for a rivalrous good.*

Why wouldn't individuals supply the level of output that equates marginal cost with the sum of individual marginal benefits as they would in a market for a rivalrous good? Return to panel (b) of Figure 5.1. There, the supply schedule for the good, labeled S, indicates the social marginal cost of producing various levels of the good. If Jill, who has demand D_2, for the nonrivalrous good makes her own decision, she will purchase a quantity Q_2, where her own marginal benefit schedule crosses the supply schedule, which is less than the socially optimal quantity, Q_0, which equates social marginal benefit with social marginal cost. Jack, who has demand (D_1), would purchase Q_1 units of the good if he were the only demander in the market. But if Jack knew that Jill would make a purchase, he would not find it in his own self-interest to purchase any of the nonrivalrous good because, as the marginal benefit schedules are drawn, Jill's purchase quantity would exceed the amount that Jack would want to purchase on his own. He would have an incentive to *free-ride* on Jill's consumption, which he gets to enjoy because of nonrivalry. In other words, once Jill purchases Q_2 units, Jack would not find it in his personal interest to purchase any additional units on his own because each unit he purchases gives him less individual benefit than its price.

A market could still generate the optimal amount of the nonrivalrous good if all consumers would honestly reveal their marginal valuations. (Notice that one of the great advantages of the market for rivalrous goods is that consumers automatically reveal their marginal valuations with their purchases.) *Consumers do not normally have an incentive to reveal honestly their marginal valuations, however, when they cannot be excluded from consumption of the good.*

Congestibility: The Role of Demand

An economically relevant classification of public goods requires attention to more than their physical characteristics. At some level of demand, consumption of a good by one person may raise the marginal costs other persons face in consuming the good, so that the marginal social cost of consumption exceeds the marginal private cost. Later we will define the divergence between social and private costs as an externality. In the context of congestion, or crowding, we are dealing with a particular type of externality—an externality of consumption inflicted only on other consumers of the good.

Whether or not a particular good is congested at any particular time depends on the level of demand for the good at that time. Changes in technology, population, income, or relative prices can shift demand from levels that do not involve externalities of consumption to levels that do. For example, a road that could accommodate 1,000 vehicles per day without traffic delays might sustain substantial traffic delays in accommodating 2,000 vehicles. In some cases, seasonal or even daily shifts in demand may change the externalities of consumption, making a good more or less congested. For example, drivers may face substantial traffic delays during rush hour but no delays during midday.

We must be careful to distinguish between the marginal social cost of consumption and the marginal cost of production. A purely nonrivalrous and noncongestible good exhibits zero marginal costs of consumption. Yet increments of the good (unless they occur naturally) require various factor inputs to produce. For instance, one way

to increase the level of defense is to increase readiness, say, by shooting off more ammunition in practice. But it takes labor, machines, and materials to produce and shoot the ammunition, things that could be used to produce other goods instead. Thus, the marginal cost of production of defense is not zero; the marginal cost of consumption is likely to be zero, however, for a given level of supply.

Some care is required in thinking about congestion in cases in which supply cannot be added in arbitrarily small units. Many goods that display nonrivalry in consumption also display *lumpiness* in supply. For example, one cannot simply add small units of physical capacity to an existing bridge. To provide additional units of capacity, one must typically either build a new bridge or double-deck the existing one. Either approach provides a large addition to capacity. But this lumpiness is irrelevant to the determination of the relevance of congestion. The important consideration is whether or not the external costs of consumption are positive at the available level of supply.

To summarize, three characteristics determine the specific nature of the public good (and hence the nature of the inefficiency that would result solely from market supply): the degree of rivalry in consumption; the extent of excludability, or exclusiveness, in use; and the existence of congestion. The presence of nonrivalry, nonexcludability, or congestion arising from changes in levels of demand can lead to the failure of markets to achieve Pareto efficiency. The presence of either nonrivalry or nonexcludability is a necessary condition for the existence of a public good market failure.

A Classification of Public Goods

Figure 5.2 presents the basic taxonomy of public goods with the rivalrous/nonrivalrous distinction labeling the columns and the excludability/nonexcludability distinction labeling the rows. Additionally, the diagonals within the cells separate cases where congestion is relevant (congested) from those cases where congestion is not relevant (uncongested). By the way, note that the definitions of public goods that we provide differ starkly from a common usage of the term public good to describe any good provided by government. In fact, *governments, and indeed markets, provide both public and private goods* as defined in this taxonomy.

Rivalry, Excludability: Private Goods. The northwest (NW) cell defines private goods, characterized by both rivalry in consumption and excludability in use: shoes, books, bread, and the other things we commonly purchase and own. In the absence of congestion or other market failures, the self-interested actions of consumers and firms elicit and allocate these goods efficiently so that government intervention would have to be justified by some rationale other than the promotion of efficiency.

When the private good is congested (that is, it exhibits externalities of consumption), market supply is generally not efficient. (In competitive markets the marginal social cost of the good equals not just the price—the marginal cost of production—but the sum of price and the marginal social costs of consumption.) Rather than consider such situations as public goods market failures, it is more common to treat them under the general heading of externality market failures. Consequently, we postpone consideration of category NW2 in Figure 5.2 to our discussion of externalities.

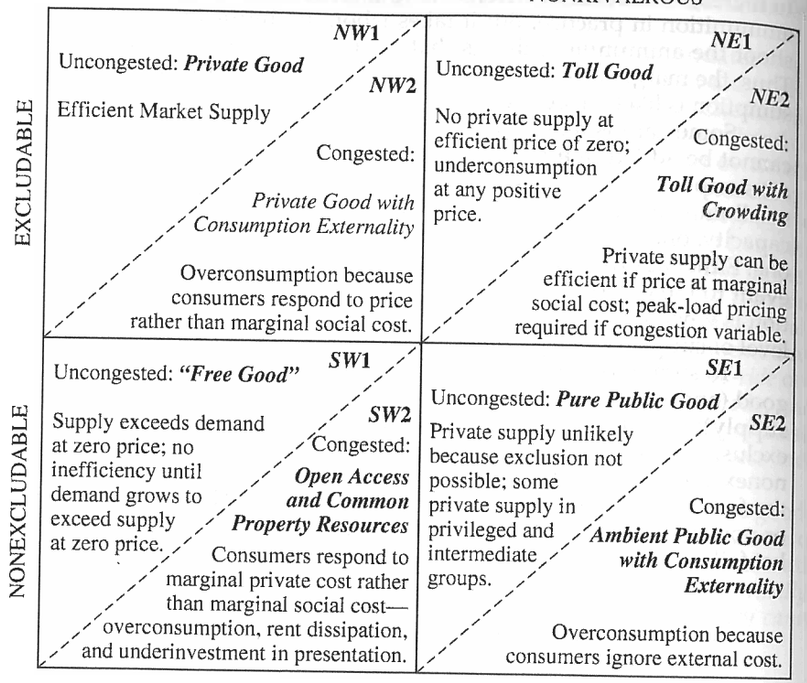


Figure 5.2 A Classification of Goods: Private and Public

Nonrivalry, Excludability: Toll Goods. The northeast cell (NE) includes those goods characterized by nonrivalry in consumption and excludability. Such goods are often referred to as *toll goods*. Prominent examples include bridges and roads that, once built, can carry significant levels of traffic without crowding. Other examples include goods that occur naturally such as wilderness areas and lakes. Because exclusion is in many cases economically feasible, a private supplier might actually come forward to provide the good. For example, an enterprising individual might decide to build a bridge and charge a crossing toll that would generate a stream of revenue more than adequate to cover the cost of construction—hence, the label toll good. Clearly, in these situations the problem is not supply, per se. Rather, the problem is twofold. First, the private supplier may not efficiently price the facility that is provided. Second, the private supplier may not provide the correct facility size to maximize social surplus [Q_S in Figure 5.1(b)].

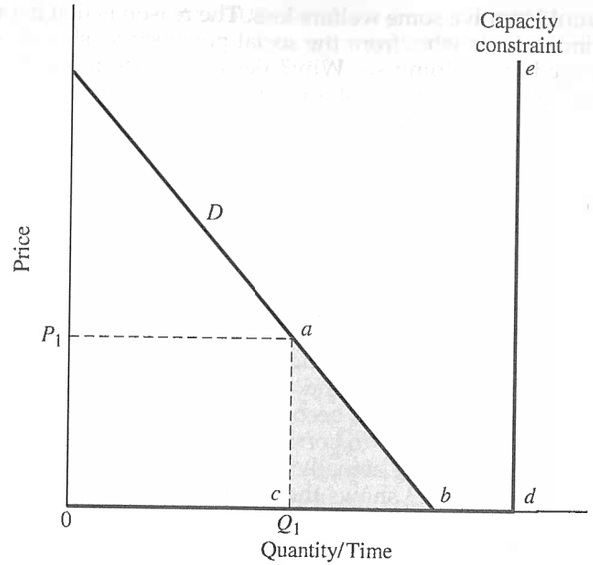
With respect to the pricing problem, first consider the case of nonrivalrousness and no congestion (NE1), one in which we can be certain that the private supplier will not charge the efficient price, which is zero. In the absence of congestion, the social marginal cost of consumption is zero, so that any positive price inappropriately discourages use of the bridge. The shaded triangular area *abc* in panel (a) of Figure 5.3 represents the deadweight loss that results from a positive price (or toll) P_1 with the demand schedule for crossings given by D. From the diagram we can see that any

positive price would involve some welfare loss. The reason is that if a positive price is charged, some individuals who, from the social perspective, should cross the bridge will be discouraged from doing so. Why? Because those individuals who would receive marginal benefits in excess of, or equal to, the marginal social cost of consumption should cross. As the marginal social cost of consumption is zero (that is, lies along the horizontal axis), any individual who would derive any positive benefit from crossing should cross. Those who obtain positive marginal benefits less than the price, however, would not choose to cross. The resulting deadweight loss may seem nebulous—the lost consumer surplus of the trips that will not take place because of the toll—but it represents a misallocation of resources that, if corrected, could lead to a Pareto improvement.

The analysis just presented is complicated if the good displays congestion over some range of demand (NE2). Return to the bridge example. We assumed that capacity exceeded demand—but this need not be the case. Consumption is typically uncongested up to some level of demand, but as additional people use the good, the marginal social costs of consumption become positive. Goods such as bridges, roads, and parks are potentially subject to congestion because of physical capacity constraints. Whether or not they are actually congested depends on demand as well as capacity. Panel (b) of Figure 5.3 shows the marginal social costs of consumption for goods that are congested. At low demand levels (for example, D_L), the marginal cost of consumption is zero (in other words, consumption is uncongested), but at high levels of demand (for example, D_H), consumption at a zero toll imposes marginal costs on all users of the good. Line segments $0f$, fg, and gh trace out the marginal social cost of consumption for the good over the range of possible consumption. Notice that these costs are imposed by the marginal consumer, not by the good itself. In the case of the bridge, for instance, the costs appear as the additional time it takes all users to cross the bridge. If additional users crowd onto the bridge, then quite conceivably marginal costs could become almost infinite: gridlock prevents anyone from crossing the bridge until some users withdraw. The economically efficient price is shown in panel (b) as P_C.

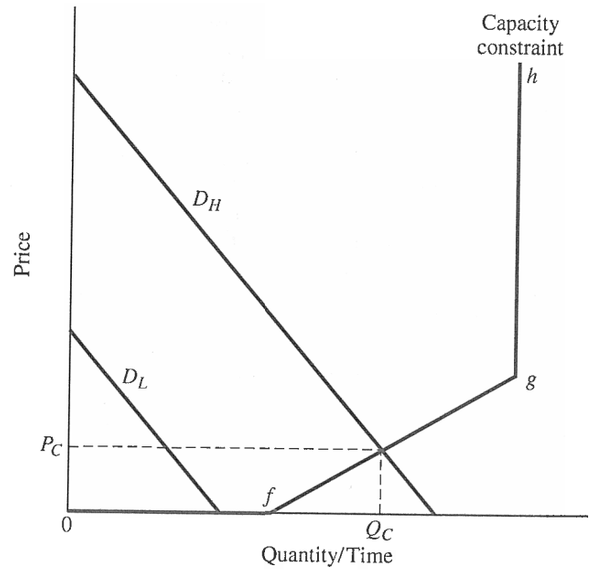
Let us take a closer look at the incentives that lead to socially inefficient crowding. Suppose that there are 999 automobiles on the bridge experiencing an average crossing time of ten minutes. You are considering whether your auto should be number 1,000 on the bridge. Given the capacity of the bridge, your trip will generate congestion—everyone crossing the bridge, including yourself, will be slowed down by one minute. In other words, the average crossing time of the users rises by one minute to 11 minutes. Your personal calculus is that you will cross if your marginal benefits from crossing exceed your marginal costs. In this case your cost is 11 minutes (the new average consumption cost of 11 minutes for the group of users). If you decide to cross, however, the marginal social costs of your decision will be 1,010 minutes (the 11 minutes you bear plus the 999 additional minutes your use inflicts on the other users). Thus, from the social perspective, if everyone places the same value on time as you, then you should cross only if the benefits that you receive from crossing exceed the cost of 1,010 minutes of delay!

In practice, nonrivalrous, excludable public goods that exhibit congestion involve quite complex pricing problems; however, the basic principle follows readily from the above discussion. Let us assume that the congestion occurs at regular time periods as demand shifts over times of the day (roads at rush hour, for instance) or seasons of the year. Efficient allocation requires that the price charged from users of the good equal the marginal costs imposed on other users during each period of the day,



Marginal social cost of consumption: Line 0*de*

(a) Social Surplus Loss with Positive Price in Absence of Congestion



Marginal social cost of consumption: Line 0*fgh*

(b) Appropriate Positive Price in Presence of Congestion

Figure 5.3 Toll Goods

implying a zero price during uncongested periods (more generally, price should equal the marginal cost of production in the absence of congestion) and some positive price during congested periods (the so-called peak-load price).

Many toll goods are produced by private firms. The firms must pay the cost of producing the goods with revenues from user fees. That is, the stream of revenues from the fees must cover the cost of construction and operation. To generate the necessary revenues, firms must usually set the tolls above the marginal social costs of consumption. Indeed, as profit maximizers, they set tolls at levels that maximize their rent [the area of rectangle P_1aQ_10 in Figure 5.3 (a)], based on the demand schedule they face. Thus, market failure results because the fees exclude users who would obtain higher marginal benefits than the marginal social costs that they impose. The magnitude of the forgone net benefits determines the seriousness of the market failure.

The problem of inefficient scale arises in the case of private supply because firms seeking to maximize profits anticipate charging tolls that restrict demand to levels that can be accommodated by smaller facilities. The result is that the facilities built are too small from a social perspective. Further, the choice of facility size determines the level of demand at which congestion becomes relevant.

Nonrivalry, Nonexcludability: Pure and Ambient Public Goods. We now turn to those goods that exhibit nonrivalrous consumption and where exclusion is not feasible—the southeast (SE) quadrant of Figure 5.2. When these goods are uncongested, they are *pure public goods*. The classic examples of such public goods are defense and lighthouses. One of the most important public goods in modern societies is the generally available stock of information that is valuable in production or consumption. With certain exceptions, to be discussed below, *pure public goods will not be supplied at all by markets* because of the inability of private providers to exclude those who do not pay for them. Contrast this with the NE quadrant, where there is likely to be market provision, but at a price that results in deadweight losses.

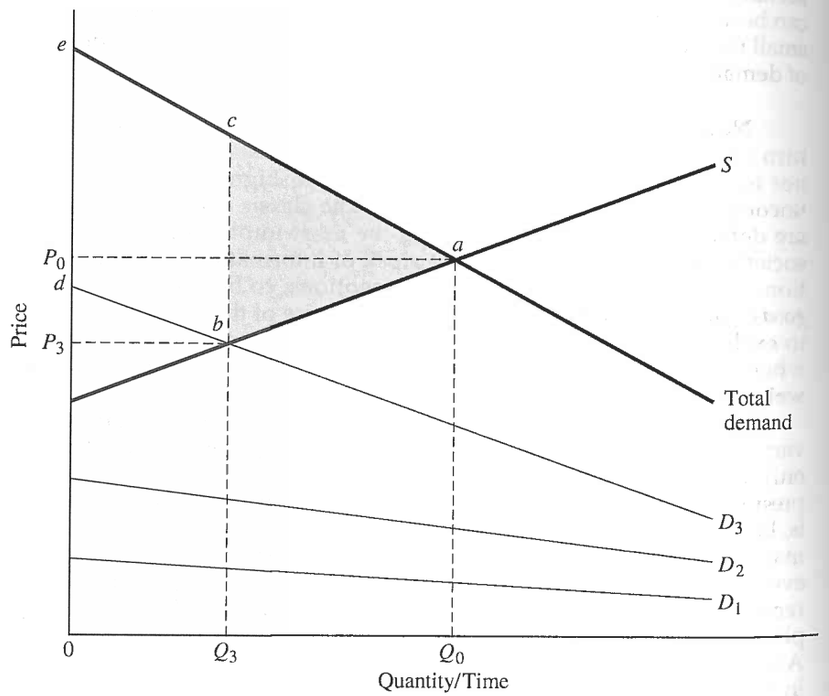
The number of persons who may potentially benefit from a pure public good can vary enormously, depending on the good: ranging from a particular streetlight, with only a few individuals benefitting to national defense, where all members of the polity presumably benefit. Because benefits normally vary spatially or geographically (that is, benefits decline monotonically as one moves away from a particular point on the map), we commonly distinguish among local, regional, national, international, and even global public goods.[5] While this is a convenient way of grouping persons who receive benefits, it is only one of many potential ways. For example, persons who place positive values on wilderness areas in the Sierras may be spread all over North America—indeed, all over the world. Some, or even most, of those who actually reside in the Sierras may not be included in this category because their private interests depend upon commercial or agricultural development of the area rather than upon preservation.

We have already touched briefly on the major problem in the SE quadrant. People who would actually receive some level of positive benefits, if the good is

[5]For an interesting discussion and classification of global public goods in health, see Todd Sandler and Daniel G. Arce, "A Conceptual Framework for Understanding Global and Transnational Public Goods for Health," *Fiscal Studies* 23(2) 2002, 195–222.

provided, often do not have an incentive to reveal honestly the magnitude of these benefits: *if contributions for a public good are to be based on benefit levels, then individuals generally have an incentive to understate their benefit levels*; if contributions are not tied to benefit levels, then individuals may have an incentive to overstate their benefit levels to obtain a larger supply. Typically, the larger number of beneficiaries, the less likely is any individual to reveal his or her preferences. In such situations, private supply is unlikely. As Mancur Olson has pointed out, however, two specific situations can result in market supply of pure public goods.[6] He labels these situations the *privileged group* and the *intermediate group* cases.

The privileged group case is illustrated in Figure 5.4, where three consumers receive benefits from the good according to their marginal benefit schedules. For example, the three might be owners of recreational facilities in a valley and the good might be spraying to control mosquitoes. The marginal benefit schedule of *person 3* (D_3) is high relative to the marginal benefit schedules of *persons 1* and 2. (By relatively high, we mean that at any quantity, *person 3* places a much higher marginal value on having an additional unit than do the other two persons.) In fact, it is sufficiently high that *person 3* will be willing to purchase Q_3 of the good if the other two persons purchase none. (*Person 3's* demand schedule intersects with the supply schedule at quantity Q_3.) Of course, once *person 3* purchases Q_3, neither *person 1* nor *person 2* will be willing to make additional purchases. In effect, they free-ride on *person 3's* high demand. (We can speak of *person 3's* marginal benefit schedule as a demand schedule because she will act as if the good were private, revealing her demand at various prices.) Despite the provision of Q_3 units of the public good, compared to the socially efficient level Q_0, social surplus is lower by the area of triangle *abc*.



Socially optimal level of provision: Q_0
Private provision by person 3: Q_3
Deadweight loss from underprovision: *abc*

Figure 5.4 Private Provision of a Public Good: Privileged Group

[6]Mancur Olson, *The Logic of Collective Action* (Cambridge, MA: Harvard University Press, 1973), 43–52.

In this case the demand of *person 3* makes up such a large fraction of total demand that the amount purchased (Q_3) is fairly close to the economically efficient level (Q_0). In this sense, the three persons form a privileged group. Even when no one person has a sufficiently high demand to make a group privileged, however, *some* provision of the good may result if the group is sufficiently small so that members can negotiate directly among themselves. We recognize such a group as "intermediate" between privileged and unprivileged: two or more members may, for whatever reasons, voluntarily join together to provide some of the good, although usually at less than the economically efficient level. Intermediate groups are generally small, or at least have a small number of members who account for a large fraction of total demand.

The situation just described closely resembles a positive externality (benefits accruing to third parties to market transactions), which we discuss in detail in the next section. Clearly, if *person 1* and *person 2* did not agree to finance jointly the public good at the efficient level (Q_0), they would nevertheless receive benefits from Q_3, the amount purchased by *person 3*. Through the purchase of Q_3, *person 3* receives the private consumer surplus benefit given by the area of triangle P_3bd and confers an external (to herself) consumer surplus benefit of area *bced* on the other group members. For analytical convenience, however, we maintain a distinction between public goods and positive externalities. We restrict the definition of an externality to those situations where provision of a good necessarily requires the joint production of a private good and a public good. We reserve the public good classification for those cases where there is no joint production. So, for example, mosquito control poses a public good problem, while chemical waste produced as a by-product of manufacturing poses an externality problem.

In many situations large numbers of individuals would benefit from the provision of a public good where no small group receives a disproportionate share of total benefits. Such cases of large numbers raise the free-rider problem with a vengeance. In situations involving large numbers, each person's demand is extremely small relative to total demand and to the cost of provision. The rational individual compares his individual marginal benefits and costs. Take as an example national defense. The logic is likely to be as follows: my monetary contribution to the financing of national defense will be infinitesimal; therefore, if I do not contribute and everyone else does, the level of defense provided, from which I cannot be effectively excluded, will be essentially the same as if I did contribute. On the other hand, if I contribute and others do not, national defense will not be provided anyway. Either way I am better off not contributing. (As we will discuss shortly, free-riding arises in other contexts besides the market provision of public goods; it can also occur when attempts are made to

supply the economically efficient quantity of the public good through public sector mechanisms.)

To summarize, then, the free-rider problem exists in the large-numbers case because it is usually impossible to get persons to reveal their true demand (marginal benefit) schedules for the good (it is even difficult to talk of individual demand schedules in this context because they are not generally observable). Even though all would potentially benefit if all persons agreed to contribute to the financing of the good so that their average contributions (in effect, the price they each paid per unit supplied) just equaled their marginal benefits, self-interest in terms of personal costs and benefits discourages honest participation.

The concept of free-riding plays an important role in the theory of public goods. There has been considerable debate among economists about the practical significance of free-riding. Both observational and experimental evidence, however, suggest that free-riding occurs, but perhaps not to the level predicted by theory.[7] John Ledyard, in a summary of the experimental literature, concludes: "(1) In one-shot trials and in the initial stages of finitely repeated trials, subjects generally provide contributions halfway between the Pareto-efficient level and the free-riding level, (2) contributions decline with repetition, and (3) face-to-face communication improves the rate of contribution."[8] Another survey concludes that higher individual marginal per capita returns from the public good raises contributions, but free-riding problems tend to be worse in larger groups than in smaller groups.[9] In the context of relatively small groups that permit face-to-face contact, such as neighborhood associations, social pressure may be brought to bear, ensuring that failure to contribute is an unpleasant experience.[10] More generally, we would expect a variety of voluntary organizations with less individual anonymity to arise to combat free-riding. There has also been considerable theoretical interest in pricing mechanisms that encourage people to reveal their preferences truthfully, though they have as yet been little used in practice.[11]

Thus far we have not considered the issue of congestion in the SE quadrant in Figure 5.2. Some goods are simply not congestible and can be placed to the left of the diagonal (SE1) within the SE quadrant. For instance, mosquito control, a local public good, involves nonexcludability, nonrivalry in consumption, and noncongestibility; no matter how many people are added to the area, the effectiveness of the eradication remains the same. Similarly, national defense, a national or international public good, in general is not subject to congestion. In contrast, nature lovers may experience disutility when they meet more than a few other hikers in a wilderness area.

We label nonrivalrous and nonexcludable goods that exhibit congestion, and thus are appropriately placed into SE2, as *ambient public goods with consumption externalities*. Air and large bodies of water are prime examples of public goods that are exogenously provided by nature. For all practical purposes, consumption (use) of these goods is nonrivalrous—more than one person can use the same unit for such purposes as disposing of pollutants.[12] In other words, consumption of the resource (via pollution) typically imposes no Pareto-relevant impact until some threshold, or ambient-carrying capacity, has been crossed. Exclusion in many market contexts is impossible, or at least extremely costly, because access to the resources is possible from many different locations. For example, pollutants can be discharged into an air shed from any location under it (or even upwind of it!).

Relatively few goods fall into category SE2. The reason is that as congestion sets in, it often becomes economically feasible to exclude users so that goods that might otherwise be in SE2 are better placed in NE2 instead. For example, to return to wilderness hiking: once the density of hikers becomes very large, it may become economically feasible for a private owner to issue passes that can be enforced by spot checks. Wilderness in this context is an ambient public good only over the range of use between the onset of crowding and the reaching of a density where the pass system becomes economically feasible. The efficiency problems associated with many of the goods that do fall into category SE2 can alternatively be viewed as market failures due to externalities. So, for example, the carrying capacity of a body of water can be viewed as an ambient public good that suffers from overconsumption. Alternatively, and more conventionally, the pollution that the body of water receives can be viewed as an externality of some other good that is overproduced.

Rivalry, Nonexcludability: Open Access, Common Property Resources, and Free Goods. Consider goods in the SW quadrant, where consumption is rivalrous, but where exclusion is not economically feasible; in other words, there is *open access* to the good. We should stress that in this quadrant we are dealing with goods that are rivalrous in consumption. Trees, fish, bison, oil, and pastureland are all rivalrous in consumption: if I take the hide from a bison, for instance, that particular hide is no longer available for you to take. Specifically, open access means that anyone who can seize units of the good controls their use. We normally use the term *open access* to describe unrestricted entry of new users. However, the same kind of property right problem arises with unrestricted use by a fixed number of individuals who already have access. Yet, in these circumstances it is often plausible to talk in terms of ownership in the sense that a fixed number of users may share collective property rights to the good. These owners may still engage in *open effort*. Nonetheless, holding the property rights gives them an incentive and possibly the means to reduce or eliminate overconsumption.

No immediate market failure appears in those cases in which the good is naturally occurring and where supply exceeds demand at zero price. As anyone can take these goods without interfering with anyone else's use, we refer to them as *free goods* (SW1). Thus, although they are theoretically rivalrous in consumption, from an efficiency perspective they are not because of the excess supply.

[7]See, for example, Linda Goetz, T. F. Glover, and B. Biswas, "The Effects of Group Size and Income on Contributions to the Corporation for Public Broadcasting," *Public Choice* 77(20) 1993, 407–14. For a review of the experimental evidence, see Robert C. Mitchell and Richard T. Carson, *Using Surveys to Value Public Goods: The Contingent Valuation Method* (Washington, DC: Resources for the Future, 1989), 133–49.

[8]John Ledyard, "Public Goods: A Survey of Experimental Research," in John H. Kagel and Alvin E. Roth, eds., *The Handbook of Experimental Economics* (Princeton, NJ: Princeton University Press, 1995), 111–94 at 121. For "real-world" evidence, see Marco Haan and Peter Kooreman, "Free Riding and the Provision of Candy Bars," *Journal of Public Economics* 83(2) 2002, 277–91.

[9]Douglas D. Davis and Charles A. Holt, *Experimental Economics* (Princeton, NJ: Princeton University Press, 1993), 332–33.

[10]Thomas S. McCaleb and Richard E. Wagner, "The Experimental Search for Free Riders: Some Reflections and Observations," *Public Choice*, 47(3) 1985, 479–90.

[11]For an overview of demand revelation mechanisms, see Dennis C. Mueller, *Public Choice II* (New York: Cambridge University Press, 1995), 124–34.

[12]Other sorts of consumption may be rivalrous. For example, taking water from a river for irrigation is rivalrous. If congested, then the water should be treated as a common property resource (SW2 in Figure 5.2) rather than as an ambient public good. The same river, however, might be nonrivalrous with respect to its capacity to carry away wastes. See Robert H. Haveman, "Common Property, Congestion, and Environmental Pollution," *Quarterly Journal of Economics* 87(2) 1973, 278–87.

In situations in which demand is higher so that there is no excess supply at zero price, we have what is usually referred to as an *open-access resource* (SW2). When access is limited to a defined group of potential users (in other words, entry is restricted), and the users own the good in common, there is *common property ownership*. The distinction between open-access and common property situations requires some elaboration. In the case of common property, the limiting of access to a defined set of persons opens the possibility of self-governance among them that reduces or eliminates open-effort inefficiencies.[13] In the case of open access, however, the threat of new entrants effectively eliminates the possibility of self-governance. Even in cases of common property, individually rational behavior by members of the defined group can lead to inefficiency in a way that may end up being indistinguishable from open access—in such cases common property results in a *common property resource problem*. Consequently, although much of the following discussion is framed in terms of open access, it is generally relevant to common property and open effort as well (recall this is the problem illustrated in Chapter 1).

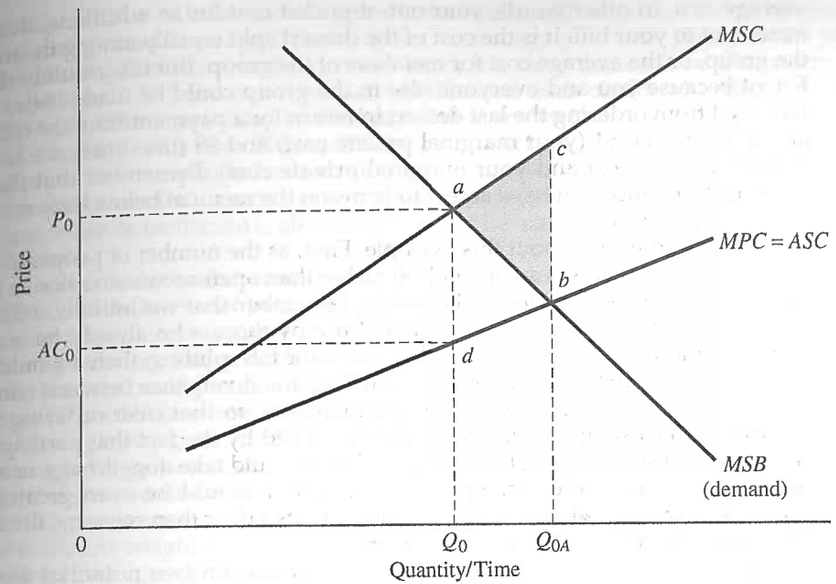
Market failure arises in the open-access case from the "infeasibility" of exclusion. Why the quotation marks? Because, as we will see, open-access problems often occur in situations in which institutional features rather than the inherent nature of the goods make exclusion infeasible. For example, in most countries oil does not suffer from open access because their governments have adopted laws that keep and enforce exclusive property rights to subsurface resources for the governments themselves. In the United States, however, oil has often suffered from the open access problem because of the "rule of capture," a legal doctrine that gives most subsurface rights to the owner of the surface property. When different people own separate pieces of property over the same pool of oil, the rule of capture prevents exclusion. Unless all the owners agree to treat their combined property as a unit, the common reservoir of oil will be extracted too quickly.

Nonexcludability leads to economically inefficient overconsumption of rivalrous goods. It can also lead to underinvestment in preserving the stock of goods or to overinvestment in capital used to capture the good.[14] Naturally occurring resources are especially susceptible to the open-access problem. Persons with access to the resource realize that what they do not consume will be consumed by someone else. Each person, therefore, has an incentive to consume the resource at a faster rate than if he or she had exclusive ownership. For instance, deforestation often results when a population relies on a forest as a source of firewood. Further, the availability of underpriced firewood does not give users the appropriate incentive to invest in stoves that use less wood, and the fact that anyone can cut and gather wood discourages individuals from replanting or nurturing trees.

Figure 5.5 illustrates the efficiency losses associated with overconsumption when there is open access to a resource. The marginal social benefit schedule *(MSB)* represents the horizontal summation (remember, we are dealing with a rivalrous good) of all the demand schedules of individuals. The economically efficient level of consumption, Q_0, results when marginal social cost *(MSC)* equals marginal social benefit *(MSC = MSB)*. Each individual, however, takes account of

[13]Gary D. Libecap, *Contracting for Property Rights* (New York: Cambridge University Press, 1989); and Elinor Ostrom, *Governing the Commons: The Evolution of Institutions for Collective Action* (New York: Cambridge University Press, 1990).

[14]See Michael B. Wallace, "Managing Resources That Are Common Property: From Kathmandu to Capitol Hill," *Journal of Policy Analysis and Management* 2(2) 1983, 220–37.



MPC = Marginal private cost
MSC = Marginal social cost
ASC = Average social cost

Socially optimal level of consumption: Q_0 ($MSB = MSC$)
Open access resource level of consumption: Q_{0A}
Deadweight loss from overconsumption: *abc*

Figure 5.5 Overconsumption of Open Access Resources

only the costs that he or she directly bears, that is, marginal private costs *(MPC)*. This private marginal cost turns out to be the average cost for all demanders *(ASC)* if marginal social cost is borne equally (that is, averaged) among consumers. With everyone rationally treating average group cost as their marginal cost, equilibrium consumption will be at Q_{OA}, which is greater than the economically efficient level Q_0. The shaded triangle *abc* measures the loss in social surplus that results from the overconsumption.

An example may help clarify why individuals in open-access and open-effort (common property) situations have an incentive to respond to marginal private cost rather than marginal social cost. Imagine that you are in a restaurant with a group of ten people who have agreed to split the bill evenly. If you were paying your own tab, then you would not order the fancy dessert costing $10 unless you expected to get at least $10 worth of value from eating it. But because the actual cost to you of the dessert will be $1 (the average increase in your bill, and for the bills of everyone else in the group—$10 divided by ten people), you would be rational (ignoring calories, your remaining stomach capacity, and social pressure) to order it as long as it would give you at least one more dollar in value. You might continue ordering desserts until the value you placed on one more fell to $1, your marginal private cost and the group

average cost. In other words, your out-of-pocket cost for an additional dessert is the increment to your bill. It is the cost of the dessert split equally among the members of the group, or the average cost for members of the group. But this result is clearly inefficient because you and everyone else in the group could be made better off if you refrained from ordering the last dessert in return for a payment from the others of any amount between $1 (your marginal private cost) and $9 (the difference between the marginal social cost and your marginal private cost). Remember that the problem arises here because you have access to items on the menu at below their real social (in this case, group) cost.

Note two things about this example. First, as the number of people in the group is restricted to ten, this is an open-effort rather than open-access situation in that membership in the group is closed. However, remember that we initially defined open-access broadly as including "unrestricted use by those who already have access." If anyone could freely join the group and join in the tab splitting, then it would be a pure open-access situation. As group size increases, the divergence between marginal private and marginal group (social) costs increases, so that over-ordering increases. Second, in this example, overordering is mitigated by the fact that participants must consume the food at the table. If the participants could take doggie bags, or even resell ordered food, then the incentives to over order would be even greater. Natural resource extractors, who typically sell seized units rather than consume them, are thus like those who are able to use doggie bags.

Modeling common property as a game between two potential users makes clear how individual incentives lead to overexploitation from a social perspective. Imagine that two ranchers have access to a grass land. Each must decide how many head of cattle to graze on the grassland not knowing how many head the other will graze. For simplicity, imagine that the choices are either 50 or 100 head of cattle. These choices are the *strategies* available to the ranchers in this game. Each pair of strategies, shown as the cells in Figure 5.6, yields pairs of *payoffs* to the two ranchers given as (payoff to *rancher 1*, payoff to *rancher 2*). If each chooses to graze 50 head, then each earns a profit of $1,000 because the total herd size of 100 can be accommodated by the field. If each chooses to graze 100 head, then each earns a profit of only $700 because the total herd size is much too large for the field so that the cattle gain too little weight. If *rancher 1* grazes 100 head and *rancher 2* grazes 50 head, then *rancher 1* earns a profit of $1,200 and *rancher 2* a profit of only $600. If *rancher 1* grazes 50 and *rancher 2* grazes 100, then it is *rancher 2* who earns the $1,200 and *rancher 1* the $600. In these strategy combinations, the herd size is too big for the field, but the lost weight per head does not offset the advantage to the more aggressive rancher of grazing the extra 50 head.

		Rancher 2	
		50 Head	100 Head
Rancher 1	50 Head	($1,000, $1,000)	($600, $1,200)
	100 Head	($1,200, $600)	($700, $700)

Figure 5.6 Choice of Herd Size as a Prisoner's Dilemma

A prediction of behavior in a game is the *Nash equilibrium. In a two-person game, a pair of strategies is a Nash equilibrium if, given the strategy of the other player, neither player wishes to change strategies.* In the game at hand, it is clear that each rancher restricting his herd size to 50 head is not a Nash equilibrium: each could raise his profit from $1,000 to $1,200 by switching to 100 head. But if one switched to 100 head, the other could raise his profits from $600 to $700 by also switching. Only when both choose 100 head would neither player have an incentive to move back to 50 head. Thus, the only Nash equilibrium in this game is for both to choose herds of 100 head. That this equilibrium is Pareto inefficient is clear—each would be made better off if they chose herds of 50 head.

Games with similar structure, called *prisoner's dilemmas*, are widely used by social scientists to model problems of cooperation.[15] They are called *noncooperative games* because of the assumption that the players cannot make binding commitments to strategies before they must be chosen. If it were possible to make binding commitments, then we could imagine the ranchers cooperating by agreeing to limit their herd sizes to 50 head before strategies are chosen. As this game is only played one time, it is called a *single-play game*. One could imagine that the ranchers faced the problem of choosing herd sizes each year. It might then be reasonable to model their interaction as a *repeated game* consisting of successive plays of the single-play, or *stage*, game. If the repeated game consists of an infinite, or uncertain, number of repetitions, and players give sufficient weight to future payoffs relative to current payoffs, then cooperative equilibria may emerge that involve each repeatedly choosing strategies that would not be equilibria in the stage game. (We return to these ideas when we discuss corporate culture and leadership in Chapter 12.)

Natural resources (both renewable and nonrenewable) have the potential for yielding scarcity rents, or returns in excess of the cost of production. With open access or open effort, these rents may be completely dissipated. In Figure 5.5, for instance, consumption at the economically efficient level Q_0 would yield rent equal to the area of rectangle P_0adAC_0. The economically efficient harvesting of salmon, for example, requires catch limits that keep the market price above the marginal costs of harvesting. In the absence of exclusion, however, these rents may well be completely dissipated.[16] (At consumption level Q_{OA} in Figure 5.5, there is no rent.) The reason is that fishers will continue to enter the industry (and those already in it will fish more intensively) as long as marginal private benefits, the rents they can capture, exceed the marginal private costs. Just as in the restaurant example, each fisher will ignore the marginal costs that his behavior imposes on other fishers. If every fisher is equally efficient, then his marginal private cost equals the average cost for the fishers as a group.

The question of how rent should be distributed is usually one of the most contentious issues in public policy. For example, how should the catch limits for salmon be divided among commercial, sport, and Native fishers? Nevertheless, from the perspective of economic efficiency someone should receive the scarcity rent rather than allowing it to be wasted. Indeed, economic efficiency as measured by reductions in the

[15]For introductions to game theory, see James D. Morrow, *Game Theory for Political Scientists* (Princeton, NJ: Princeton University Press, 1994); and Martin J. Osborne and Ariel Rubenstein, *A Course in Game Theory* (Cambridge, MA: MIT Press, 1994).

[16]See, for example, L. G. Anderson, *The Economics of Fishery Management* (Baltimore, MD: Johns Hopkins University Press, 1977).

dissipation of rent was one of the goals of the regulatory alternatives for the British Columbia salmon fishery analyzed in Chapter 1.

Note that demand for a resource may be sufficiently low so that it remains uncongested at zero price. Increases in demand, however, may push the resource from a free good (SW1) to an open-access good (SW2). A number of free goods, including such resources as buffalo, forests, aquifer water, and rangeland, have been historically important in North America. Typically, however, the free goods eventually disappeared as demand increased and excess supply was eliminated.[17] Open access often led to rapid depletion and, in some cases, near destruction of the resource before effective exclusion was achieved. For example, open access permitted destruction of the Michigan, Wisconsin, and Minnesota pine forests.[18] Nonexcludability continues to be at the heart of many water use problems in the western United States.[19] When animal populations are treated as open-access resources, the final result of overconsumption may be species extinction; such a result has already occurred with certain birds and other animals with valuable organs or fur.

Thus far we have not specified the meaning of "feasible" exclusion. Many of the goods we have given as examples appear not to be inherently nonexcludable. Indeed, one of the most famous historical examples of open-access resource—sheep and cattle grazing on the medieval English pasture—was "solved," willy-nilly, without overt government intervention, by the enclosure movement, which secured property rights for estate holders. Similar enclosures appear to be currently taking place on tribal, historically open-access, lands in parts of Africa.

It is useful to dichotomize open-access and common property problems into those that are *structural* (where aspects of the goods preclude economically feasible exclusion mechanisms) and those that are *institutional* (where economically efficient exclusion mechanisms are feasible but the distribution of property rights precludes their implementation). The averaging of restaurant bills, which we previously discussed, serves as an excellent illustration of an institutional common property problem. We can imagine an obvious exclusion mechanism that we know from experience is economically feasible: separate bills for everyone. Institutional common property resource problems are usually not fundamentally market failures. Rather, they are most often due to the failure of government to allocate enforceable property rights (again the type of situation explored in the salmon fishery example in Chapter 1).

Typically, the crucial factor in making a distinction between structural and institutional problems is whether or not the good displays *spatial stationarity*. Trees are spatially stationary, salmon are not, and bodies of water may or may not be. When resources are spatially stationary, their ownership can be attached to the ownership of land. Owners of the land are usually able to monitor effectively all aspects of their property rights and, consequently, ensure exclusive use. Given exclusion, common property resources become private resources that will be used in an economically

[17]In the case of the buffalo, the opening of the railroad facilitated the hunting and transportation of hides at much lower costs, so that what had previously been a free good became an open-access resource. See John Hanner, "Government Response to the Buffalo Hide Trade, 1871–1883," *Journal of Law and Economics* 24(2) 1981, 239–71.

[18]Andrew Dana and John Baden, "The New Resource Economics: Toward an Ideological Synthesis," *Policy Studies Journal* 14(2) 1985, 233–43.

[19]B. Delworth Gardner, "Institutional Impediments to Efficient Water Allocation," *Policy Studies Review* 5(2) 1985, 353–63; William Blomquist and Elinor Ostrom, "Institutional Capacity and the Resolution of a Commons Dilemma," *Policy Studies Review* 5(2) 1985, 283–93.

efficient manner. Without spatial stationarity, ownership of land is not a good proxy for low monitoring costs and the viability of enforcing exclusion. It does not necessarily follow that the open-access or common property problem could not be dealt with by some form of private ownership, but it does suggest that ownership of a defined piece of land or water will not be adequate to ensure exclusion. Allocating fishing rights to specific water acreage where the fish stock moves over considerable distances, or associating the rights to oil extraction to ownership of land when the oil pool extends under a large number of parcels, illustrate the difficulty of creating effective property rights for nonstationary resources.

In summary, a stationary good may have common property resource characteristics simply because its ownership is not well defined, perhaps because of the historical accident that at one time supply exceeded demand at zero price. Nonstationary goods generally require more complex policy interventions to achieve efficiency because the linking of property rights to landownership will not serve as an effective proxy for exclusive resource ownership of the resource.

Reprise of Public Goods

Returning to Figure 5.2, we summarize the efficiency implications of the various types of market failures involving public goods. To reprise, the major problem with toll goods (NE quadrant—nonrivalry, excludability) is underconsumption arising from economically inefficient pricing rather than a lack of supply per se. Congestion usually further complicates these problems by introducing the need for variable pricing to achieve efficiency. In the case of pure and ambient public goods (SE quadrant—nonrivalry, nonexcludability), the pervasiveness of free-riding generally leads to no market supply at all. In specific circumstances (a privileged or intermediate group in which one or a few persons account for a large fraction of demand), however, there may be some, and perhaps even nearly efficient, market supply. In the case of open access resources (SW quadrant—rivalry, nonexcludability), inefficiency results because individuals do not equate marginal social costs with marginal benefits but, rather, marginal private costs with marginal benefits. Hence, they inefficiently overconsume and inefficiently underinvest in enhancing open-access resources.

Externalities

An *externality* is any valued impact (positive or negative) resulting from any action (whether related to production or consumption) that affects someone who did not fully consent to it through participation in voluntary exchange. Price changes in competitive markets are not relevant externalities because buyers and sellers are engaging voluntarily in exchange. We have already encountered a variety of externalities in our discussion of public goods: private supply of nonrivalrous goods by privileged and intermediate groups (a positive externality) and the divergence between marginal private and marginal social costs in the use of congested resources (a negative externality). We reserve the label *externality problem* for those situations in which the good conveying the valued impact on nonconsenting parties is the by-product of either the production or consumption of some good.

As is the case with open-access resources and ambient public goods, externality problems involve attenuated property rights because either the rights to exclusive use are incompletely specified or the costs of enforcing the rights are high relative to the benefits. Secure and enforceable property rights often permit private transactions to eliminate the economic inefficiency associated with an externality by opening up the possibility for markets in the external effect. Indeed, one can think of an externality problem as a *missing market*. We return to this point after discussing a few examples.

Common examples of negative externalities include the air and water pollution generated by firms in their production activities, the cigarette smoke that nonsmokers must still breathe in public places in some countries, and the unsightliness generated by a dilapidated house in a well-kept neighborhood. Persons who suffer these externalities place different values on them. For instance, you may really fear the health consequences of second-hand cigarette smoke, but I may be too old to worry about it. Whereas I would be willing to pay only a small cost to avoid sitting near a smoker, say, waiting an extra ten minutes for a restaurant table in the nonsmoking section, you might be willing to pay considerably more, say, waiting thirty minutes or leaving the restaurant altogether. Note that we can think of placing a value on these externalities in the same way we do for the goods we voluntarily consume.

Common examples of positive externalities include vaccinations that reduce everyone's risk of infectious disease (so-called herd immunity) and the benefits that neighbors receive from a homeowner's flower garden and nicely painted house. An important category of positive externality arises in the context of communication networks. One more person connecting to a Web-based digital music exchange provides marginal social benefits that exceed marginal private benefits because everyone already on the network has one more interface for potential exchange.[20] Such positive externalities are usually referred to as *network*, or *adoption, externalities*.

Externalities can arise in either production or consumption. Production externalities affect either firms (producer-to-producer externalities) or consumers (producer-to-consumer externalities); consumption externalities may also affect the activities of firms (consumer-to-producer externalities) or those of other consumers (consumer-to-consumer externalities). In this context, the category of consumers that are the recipients of externalities includes everyone in society. Table 5.1 provides simple examples of each type of externality. (Keep in mind that sometimes the same activity may constitute a positive externality for some but a negative externality for others.) Classifying a situation that potentially involves an externality is often a good way to begin considering its efficiency implications, distributional impacts, and, most important, possible remedies.

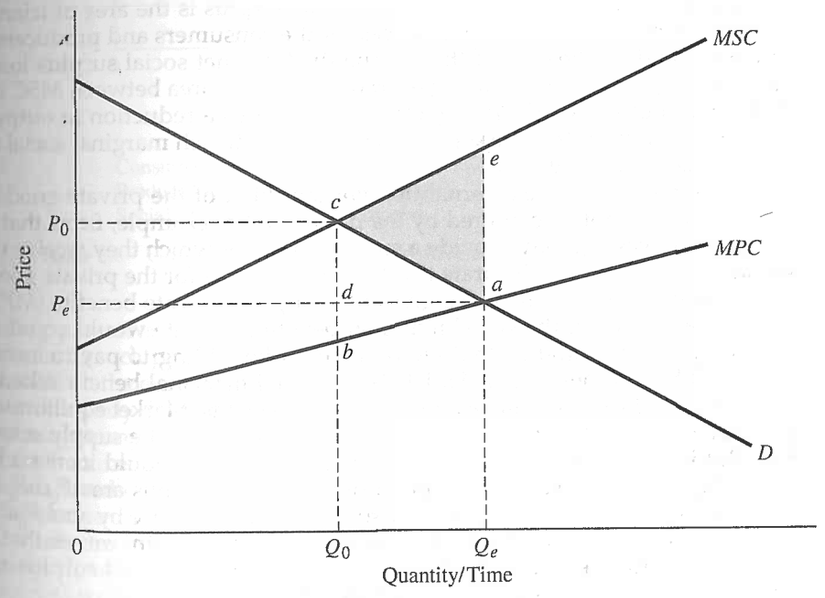
Efficiency Losses of Negative and Positive Externalities

Figure 5.7 illustrates the resource allocation effects of a negative externality in production. In the presence of a negative externality, firms will produce too much of the private good that generates the externality. The market supply schedule for the private

[20]For an analysis of network externalities, see Hal R. Varian, *Intermediate Economics: A Modern Approach* (New York: W.W. Norton and Co., 1999), 606–12.

Table 5.1 *Examples of Externalities*

	Positive	Negative
Producer-to-Producer	Recreational facilities attracting people who give custom to nearby businesses	Toxic chemical pollution harming downstream commercial fishing
Producer-to-Consumer	Private timber forests providing scenic benefits to nature lovers	Air pollution from factories harming lungs of people living nearby
Consumer-to-Consumer	Immunization of persons against contagious disease helping reduce risk to others	Cigarette smoke from one person reducing enjoyment of meal by another
Consumer-to-Producer	Unsolicited letters from consumers providing information on product quality	Game hunters disturbing domestic farm animals



Social Surplus at Q_e relative to Q_0

Consumer surplus from private good:	larger by P_0caP_e
Producer surplus of firms producing private good:	smaller by $(P_0cdP_e - abd)$
Losses to third parties bearing externality:	larger by $abce$
Social surplus:	smaller by ace

Figure 5.7 Overproduction with a Negative Externality

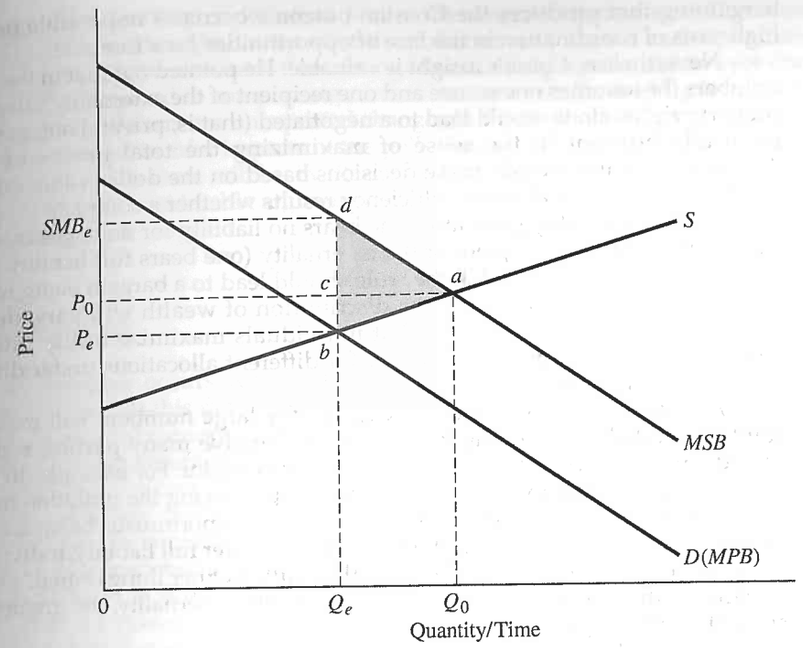
good, *MPC*, indicates the marginal costs borne directly by the firms producing it. For example, if the private good were electricity, *MPC* would represent the marginal amounts firms have to pay for the coal, labor, and other things that show up in their ledgers. But *MPC* does not reflect the negative impacts of the pollution that results from burning the coal. If somehow we could find out how much each person in society would be willing to pay to avoid the pollution at each output level, then we could add these amounts to the marginal costs actually seen by the firms to derive a supply schedule that reflected total social marginal costs. We represent this more inclusive supply schedule as *MSC*.

Economic efficiency requires that social marginal benefits and social marginal costs be equal at the selected output level; this occurs at quantity Q_0, where marginal social cost *(MSC)* and demand *(D)* intersect. But because firms do not consider the external costs of their output, they choose output level Q_e at the intersection of *MPC* and *D*. Relative to output level Q_0, consumers of the private good being produced gain surplus equal to area P_0caP_e (because they get quantity Q_e at price P_e rather than quantity Q_0 at price P_0) and producers lose surplus equal to area P_0cdP_e minus area *abd* (the first area captures the effect of the lower price, the second the effect of greater output). Those who bear the external costs, however, suffer a loss given by the area *abce* (the area between the market and social supply curves over the output difference; remember, the vertical distance between the supply curves represents the external marginal cost). The net loss in social surplus is the area of triangle *ace*, the algebraic sum of the surplus differences for the consumers and producers of the private good and the bearers of the externality. (This net social surplus loss is simply the deadweight loss due to the overproduction—the area between *MSC* and *D* from Q_0 to Q_e.) In other words, Pareto efficiency requires a reduction in output from the equilibrium level in the market (Q_e) to the level at which marginal social costs equal marginal social benefits (Q_0).

Turning to positive externalities, we can think of the private good generating benefits that cannot be captured by the producer. For example, firms that plant trees for future harvesting may provide a scenic benefit for which they receive no compensation. In Figure 5.8 we illustrate the demand schedule for the private good (trees for future harvest) as *D*, which also gives the marginal private benefits *(MPB)*. At each forest size, however, the social marginal benefit schedule would equal the market demand plus the amounts that viewers would be willing to pay to have the forest expanded by another unit. *MSB* labels the social marginal benefit schedule, which includes both the private and external marginal benefits. Market equilibrium results at output level Q_e, where the market demand schedule and the supply schedule intersect. But if output were raised to Q_0, consumer surplus would increase by area *acd* (resulting from increased consumption from Q_e to Q_0) minus area P_0cbP_e (due to the price rise from P_e to P_0), and producer surplus would increase by area P_0abP_e. The net increase in social surplus, therefore, would be area *abd*. Again, we see that in the case of an externality, we can find a reallocation that increases social surplus and thereby offers the possibility of an increase in efficiency.

Market Responses to Externalities

Will the market always fail to provide an efficient output level in the presence of externalities? Just as pure public goods will sometimes be provided at efficient, or nearly efficient, levels through voluntary private agreements within intermediate groups, so



Social Surplus at Q_e relative to Q_0

Consumer surplus:	smaller by $(acd - P_0cbP_e)$
Producer surplus:	smaller by P_0abP_e
Social surplus:	smaller by *abd*

Figure 5.8 Underproduction with a Positive Externality

too may private actions counter the inefficiency associated with externalities. The relevance of such private responses was first pointed out by Ronald Coase in a seminal article on the externality problem.[21] He argued that in situations in which property rights are clearly defined and costless to enforce, and utility is linear in wealth, costless bargaining among participants will lead to an economically efficient level of external effect. Of course, the distribution of costs and benefits resulting from the bargaining depends on who owns the property rights.

Before exploring the issue further, we must stress a very restrictive assumption of Coase's model that limits its applicability to many actual externality situations. Namely, Coase assumes zero transaction costs in the exercise of property rights.[22] In many real-world situations, transaction costs are high usually because those producing and experiencing the externality are numerous. With large numbers of actors, the

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

[22]For a conceptual discussion of transaction costs taking account of bargaining, see Douglas D. Heckathorn and Steven M. Maser, "Bargaining and the Sources of Transaction Costs: The Case of Government Regulation," *Journal of Law, Economics, and Organization* 3(1) 1987, 69–98.

bargaining that produces the Coasian outcome becomes impossible because of the high costs of coordination in the face of opportunities for a free ride.

Nevertheless, Coase's insight is valuable. He pointed out that in the case of small numbers (he assumes one source and one recipient of the externality), the allocation of property rights alone would lead to a negotiated (that is, private) outcome that is economically efficient in the sense of maximizing the total profits of the parties. Assuming that individuals make decisions based on the dollar value rather than the utility value of external effects, efficiency results whether a complete property right is given to the externality generator (one bears no liability for damages caused by one's externality) or to the recipient of the externality (one bears full liability for damages caused by one's externality). Either rule should lead to a bargain being reached at the same level of externality; only the distribution of wealth will vary, depending on which rule has force. Assuming that individuals maximize utility rather than net wealth, however, opens the possibility for different allocations under different property rights assignments.[23]

A moment's thought should suggest why large numbers will make a Coasian solution unlikely. Bargaining would have to involve many parties, some of whom would have an incentive to engage in strategic behavior. For example, in the case of a polluter with no liability for damages, those experiencing the pollution may fear free-riding by others. In addition, firms may engage in opportunistic behavior, threatening to generate more pollution to induce payments. Under full liability, individuals would have an incentive to overstate the harm they suffer. Other things equal, we expect that the greater the number of parties experiencing the externality, the greater will be the costs of monitoring damage claims.

Nevertheless, private cooperation appears effective in some situations. For instance, neighborhood associations sometimes do agree on mutually restrictive covenants, and individual neighbors occasionally do reach contractual agreements on such matters as light easements (which deal with the externalities of the shadows cast by buildings).

Moreover, there is an important case where Coase-like solutions do arise, even with large numbers of parties—namely, where (1) property rights become implicitly established by usage; (2) the value of the externality (whether positive or negative) is captured by (more technically, "capitalized into") land values; (3) considerable time has passed such that the initial stock of external parties has "rolled over"; and (4) externality levels remain stable. The relevance of these conditions can best be explained with an example. Suppose that a factory has been polluting the surrounding area for many years without anyone challenging its owner's right to do so. It is probable that the pollution will result in lower property values.[24] Residents who bought before the pollution was anticipated will have to sell their houses for less—reflecting the impact of the pollution. New homeowners, however, will not bear any Pareto-relevant externality, because the negative impact of the pollution will be capitalized into house prices. The lower prices of the houses will reflect the market's (negative) valuation of the pollution. In other words, through the house price it is possible to get a proxy dollar

[23]For a discussion of this point and an overview of Coase, see Thrainn Eggertsson, *Economic Behavior and Institutions* (New York: Cambridge University Press, 1990), 101–10.

[24]Indeed, changes in property values provide a basis for empirically estimating the social costs of externalities. For example, with respect to air pollution, see V. Kerry Smith and Ju-Chin Huang, "Can Markets Value Air Quality? A Meta-Analysis of Hedonic Property Value Models," *Journal of Political Economy* 103(1) 1995, 209–27.

measure of the disutility of pollution. Notice that a second generation of homeowners (that is, those who bought houses after the pollution was known and capitalized into prices) would receive a bonus if the existing allocation of property rights were changed so that the factory had to compensate current homeowners for existing levels of pollution. Of course, if there are unexpected changes in the level of pollution (or new information about the harmful impacts of the pollution—see our discussion of information asymmetry below), there will be, in effect, new (either positive or negative) impacts; in these situations, considerable argument is likely to occur over who has rights to compensation for the changes.

Natural Monopoly

Natural monopoly occurs when average cost declines over the relevant range of demand. Note that this definition is in terms of both cost and demand conditions. In the case of natural monopoly, a single firm can produce the output at lower cost than any other market arrangement, including competition.

Although the cost-and-demand conditions establish the existence of a natural monopoly situation, the *price elasticity of demand* determines whether or not the natural monopoly has important implications for public policy. The price elasticity of demand measures how responsive consumers are to price changes. Specifically, the price elasticity of demand is defined as the percentage change in the quantity demanded that results from a 1 percent change in price.[25] If the absolute value of the price elasticity of demand is less than one (a 1 percent change in price leads to less than a 1 percent reduction in the quantity demanded), then we say that demand is inelastic and an increase in price will increase total revenue. A good is unlikely to have inelastic demand if there are other products that, while not exactly the same, are close substitutes. In such circumstances, the availability of substitutes greatly limits the economic inefficiency associated with natural monopoly. For example, although local cable in some television markets may have the cost-and-demand characteristics of natural monopolies, many substitute products, including over-the-air television, satellite TV providers, and DVD players, may prevent cable television companies from realizing large monopoly rents in their local markets.

We should also keep in mind that, although we stated the basic definition of natural monopoly in static terms, markets in the real world are dynamic. Technological change may lead to different cost characteristics, or high prices may bring on close or even superior substitutes. The result may be the elimination of natural monopoly or the supplanting of one natural monopoly technology by another.[26] For example, especially in developing countries, cellular phones are direct substitutes for land-line phones, providing competition for traditional telephone systems.

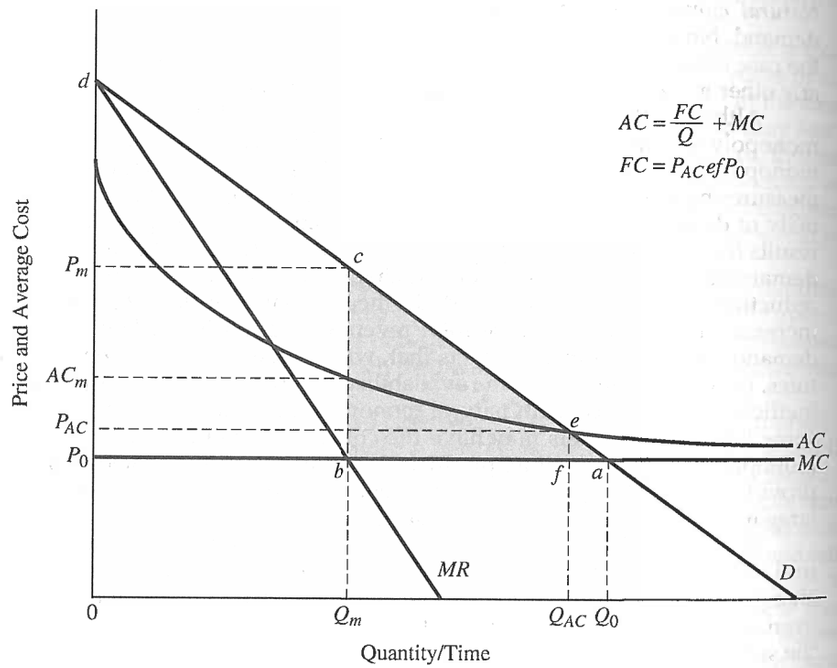
[25]Mathematically, if the demand schedule is continuous, the price elasticity of demand at some quantity equals the slope of the demand schedule at that quantity times the ratio of quantity to price. For example, with linear demand schedule, $Q = a - bP$, the slope of the demand schedule (the derivative dQ/dP) is $-b$. Therefore, the price elasticity of demand is $e = -bP/Q$. Note that the elasticity of a linear demand schedule varies with quantity.

[26]For an example of a case study that finds that technological change significantly reduced economies of scale and eliminated natural monopoly, see Stephen M. Law and James F. Nolan, "Measuring the Impact of Regulation: A Study of Canadian Basic Cable Television," *Review of Industrial Organization* 21(3) 2002, 231–49.

Allocative Inefficiency under Natural Monopoly

We show a cost structure leading to natural monopoly in Figure 5.9. Marginal cost *(MC)* is shown as constant over the range of output. Fixed cost *(FC)*, which is not shown in the figure, must be incurred before any output can be supplied. Because of the fixed cost, average cost *(AC)* starts higher than marginal cost and falls toward it as the output level increases. ($AC = FC/Q + MC$, where Q is the output level.) Although marginal cost is shown as constant, the same analysis would apply even if marginal cost were rising or falling, provided that it remains small relative to fixed cost.

Figure 5.9 shows the divergence between the firm's profit-maximizing behavior and economic efficiency. Let us first consider the economically efficient price and



	Monopoly Pricing (P_m)	Efficient Pricing (P_0)	Average Cost Pricing (P_{AC})
Consumer surplus:	P_mcd	P_0ad	$P_{AC}ed$
Total revenue:	P_mcQ_m0	P_0aQ_00	$P_{AC}eQ_{AC}0$
Total cost:	$FC + P_0bQ_m0$	$FC + P_0aQ_00$	$FC + P_0fQ_{AC}0$
Producer surplus:	$P_mcbP_0 - FC$	$-FC$	0
Social surplus:	$P_0dcb - FC$	$P_0ad - FC$	$P_{AC}ed$

Net social surplus loss of monopoly pricing relative to efficient pricing: *abc*
Net social surplus loss of average cost pricing relative to efficient pricing: *aef*

Figure 5.9 Social Surplus Loss from Natural Monopoly

output. Efficiency requires, as we discussed in the previous chapter, that price be set equal to marginal cost ($P = MC$). Because marginal cost is below-average cost, the economically efficient output level Q_0 results in the firm suffering a loss equal to *FC*. Obviously, the firm would not choose this output level on its own. Instead, it would maximize profits by selecting output level Q_m, which equates marginal revenue to marginal cost ($MR = MC$). At output level Q_m, the market price will be P_m and profits equal to the area of rectangle $P_mcbP_0 - FC$ will result. Relative to output level Q_0, consumer surplus is lower by the area of P_mcaP_0, but the profit of the firm is larger by P_mcbP_0. The net loss in social surplus due to the underproduction equals the area of the shaded triangle *abc*, the deadweight loss to consumers. (As noted in our discussion of Figure 4.5, units of forgone output would have offered marginal benefits in excess of marginal costs—a total loss equal to the area between the marginal benefit, or demand schedule, and the marginal cost curve. Again, in Figure 5.9, this loss is the area of triangle *abc*.)

Suppose that public policy forced the natural monopolist to price at the economically efficient level (P_0). The monopolist would suffer a loss of *FC*. Consequently, the monopolist would go out of business in the absence of a subsidy to offset the loss. Notice the dilemma presented by natural monopoly: *forcing the monopolist to price efficiently drives it out of business; allowing the monopolist to set price to maximize profits results in deadweight loss.*

Briefly consider what would happen if the firm were forced by public policy to price at average, rather than marginal, cost; that is, if the firm priced at P_{AC} and produced Q_{AC} in Figure 5.9. Clearly, under these circumstances the firm can survive because its costs are now being just covered. (Note that *FC* equals $P_{AC}efP_0$.) Although the deadweight loss under average cost pricing is much lower than under monopoly pricing (area *aef* versus area *abc*), it is not eliminated. Therefore, *average cost pricing represents a compromise to the natural monopoly dilemma.*

Restraints When Markets Are Contestable

We can imagine circumstances in which the natural monopoly firm might, in fact, be forced to price at average cost because of the threat of competition from potential entrants. The crucial requirement is that entry to, and exit from, the industry be relatively easy. Whether or not the firm has in-place capital that has no alternative use, capital whose costs are *sunk*, usually determines the viability of potential entry and exit. If the established natural monopoly has a large stock of productive capital that cannot be sold for use in other industries, it will be difficult for other firms to compete because they must first incur cost to catch up with the established firm's capital advantage. The in-place capital serves as a barrier to entry; the greater the replacement cost of such capital, the higher the barrier to entry. For example, once a petroleum pipeline is built between two cities, its scrap value is likely to be substantially less than the costs a potential competitor would face in building a second pipeline. The greater is the difference, the greater is the ability of the established firm to fight off an entry attempt with temporarily lower prices. Of course, keep in mind that the ability of the firm to charge above the marginal cost level will be influenced by the marginal costs of alternative transportation modes, such as truck and rail, which can serve as substitutes.

Much economic research considers industries with low barriers to entry and decreasing average costs, which, because of the threat of potential entry, are said to be

in *contestable markets*.[27] We expect markets that are contestable to exhibit pricing closer to the efficient level. One of the most important debates in the literature arising from the contestable market framework concerns the empirical significance of in-place capital as effective barriers to entry.[28] Most natural monopolies appear to enjoy large advantages in in-place capital, raising the question of whether they should be viewed as being in contestable markets.

As we have seen, the "naturalness" of a monopoly is determined by the presence of decreasing average cost over the relevant range of output. In many situations it appears that average cost declines over considerable ranges of output but then flattens out. In other words, initial economies of scale are exhausted as output increases. What happens if demand shifts to the right (for example, with population increases) such that the demand curve intersects the flat portion of the average cost curve? There may be considerable room for competition under these circumstances. Figure 5.10 illustrates such a situation. If demand is only D_1 (the "classic" natural monopoly), only one firm can survive in the market. If demand shifts outward beyond D_2 to, say, D_3, two or more firms may be able to survive because each can operate on the flat portion of the average cost curve.

Simply because two firms could survive at D_3 does not mean that two competing firms will actually emerge as demand expands. If the original natural monopoly firm expands as demand moves out, it may be able to forestall new entrants and capture all of the market. Nevertheless, we are again reminded of the importance of looking beyond the static view.

When considering natural monopoly from a policy perspective, we often find that legal and regulatory boundaries do not correspond to the boundaries delineating natural monopoly goods. Most discussion of industrial issues tends to be within the framework of product "sectors," such as the electric and the telephone industries. Unfortunately, the economic boundaries of natural monopolies are not likely to conform to these neat sectoral boundaries. Historically, regulation has often not recognized this unpleasant fact. In addition, the existence, or extent, of a natural monopoly can change as production technology or demand changes.

The telecommunications industry illustrates these definitional problems. In 1982, the U.S. Justice Department and AT&T agreed on a plan for the breakup of the corporation. The agreement called for a division into two parts: the part that could be expected to be workably contestable and the part that had strong natural monopoly elements. Both long-distance services (contestable) and equipment manufacture and supply (competitive) were largely deregulated, while local telephone exchanges were deemed to be regional natural monopolies.[29] Similar problems with sectoral

[27]William J. Baumol, John C. Panzar, and Robert D. Willig, *Contestable Markets and the Theory of Industry Structure* (New York: Harcourt Brace Jovanovich, 1982); and William J. Baumol, "Contestable Markets: An Uprising in the Theory of Industry Structure," *American Economic Review* 72(1) 1982, 1–15. For a discussion of *imperfect contestability*, see Steven A. Morrison and Clifford Winston, "'Empirical Implications and Tests of the Contestability Hypothesis," *Journal of Law and Economics* 30(1) 1987, 53–66.

[28]For evidence that the U.S. Postal Service, for example, may have few natural monopoly characteristics, see Alan L. Sorkin, *The Economics of the Postal Service* (Lexington, MA: Lexington Books, 1980), Chapter 4; and Leonard Waverman, "Pricing Principles: How Should Postal Rates Be Set?" in *Perspectives on Postal Services Issues*, Roger Sherman, ed. (Washington, DC: American Enterprise Institute, 1980), 7–26.

[29]Kenneth Robinson, "Maximizing the Public Benefits of the AT&T Breakup," *Journal of Policy Analysis and Management* 5(3) 1986, 572–97. For discussion of technological change that eroded the natural monopoly characteristics of telephone services, see Irwin Manley, *Telecommunications America: Markets without Boundaries* (Westport, CT: Quorum, 1984).

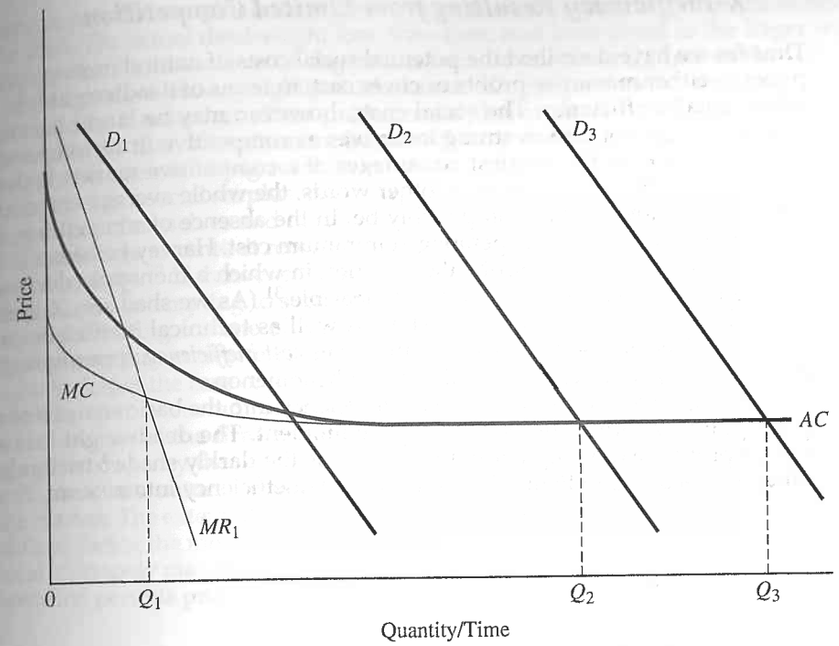


Figure 5.10 Shifting Demand and Multiple Firm Survival

definitions have not been well recognized in other contexts, however. For example, electricity generation and transmission have typically been treated as part of an electrical utility natural monopoly, although the evidence suggests that in many circumstances only transmission has the required cost and demand characteristics to be considered a natural monopoly.[30] (Running multiple sets of transmission lines across the countryside would generally be inefficient; having multiple firms generating electricity would not be.)

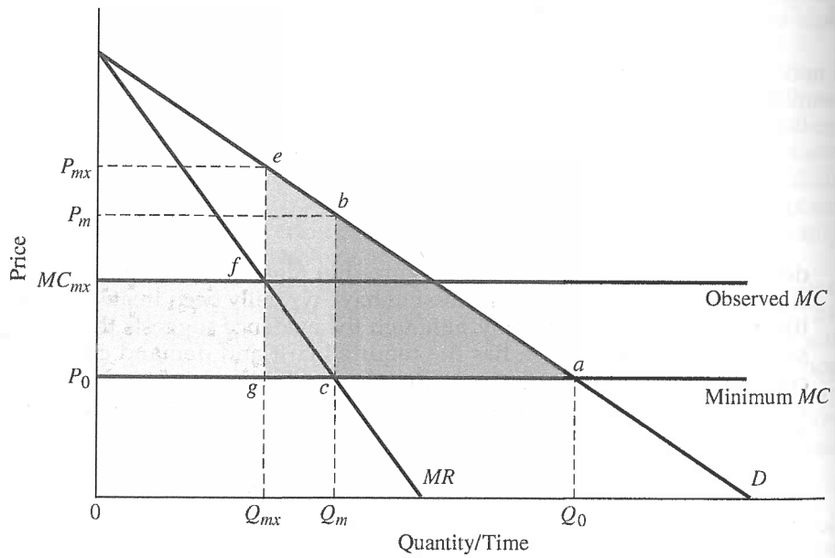
Another dimension, apart from the sectoral, where the problem of defining natural monopoly arises is the spatial. Over what spatial area does the natural monopoly exist? Almost all natural monopoly regulation corresponds to existing city, county, state, and federal boundaries; but the economic reality of a natural monopoly knows no such bounds—it is purely an empirical question how far (spatially) the natural monopoly extends. Again, the appropriate spatial boundaries of a natural monopoly can change with changes in technology.

[30]For the case against treating any aspects of the electric industry as a natural monopoly, see Robert W. Poole, Jr., *Unnatural Monopolies: The Case for Deregulating Public Utilities* (Lexington, MA: Lexington Books, 1985).

X-Inefficiency Resulting from Limited Competition

Thus far we have described the potential social costs of natural monopoly, whether it prices to either maximize profits or cover cost, in terms of deadweight loss caused by allocational inefficiency. The social costs, however, may be larger because natural monopolies do not face as strong incentives as competitive firms to operate at minimum cost. One of the greatest advantages of a competitive market is that it forces firms to keep their costs down; in other words, the whole average and marginal cost curves are as low as they can possibly be. In the absence of competition, firms may be able to survive without operating at minimum cost. Harvey Leibenstein coined the phrase *X-inefficiency* to describe the situation in which a monopoly does not achieve the minimum costs that are technically feasible.[31] (As we shall see, X-inefficiency is not fully descriptive because transfers as well as technical inefficiencies are often involved.) One also sometimes finds the terms *cost inefficiency, operating inefficiency,* or *productive inefficiency* to describe the same phenomenon.

In Figure 5.11 we incorporate X-inefficiency into the basic analysis of monopoly pricing. But ignore X-inefficiency for the moment. The deadweight loss associated with a profit-maximizing natural monopoly is the darkly shaded triangular area *abc* (already discussed in Figure 5.9). If we take X-inefficiency into account, however, the



Social surplus loss from efficient monopoly:	*abc*
Minimum social surplus loss from inefficient monopoly:	*aeg*
Maximum social surplus loss from inefficient monopoly:	$aeg + MC_{mx}fgP_0$

Figure 5.11 X-Inefficiency under Natural Monopoly

[31]See Harvey J. Leibenstein, *Beyond Economic Man* (Cambridge, MA: Harvard University Press, 1976); and Roger S. Frantz, *X-Efficiency: Theory, Evidence and Applications* (Boston, MA: Kluwer, 1988).

minimum possible marginal cost curve is lower than that observed from the behavior of the firm. The actual deadweight loss, therefore, is at least equal to the larger triangular *aeg*—the additional loss of *cbeg* results because output is Q_{mx} rather than Q_m.

Some or all of the very lightly shaded area $MC_{mx}fgP_0$ represents unnecessary cost that should be counted as either producer surplus or deadweight loss. If marginal costs are higher than the minimum because the managers of the firms employ more real resources such as hiring workers who stand idle, then the area $MC_{mx}fgP_0$ should be thought of as a social surplus loss. If, on the other hand, costs are higher because the managers pay themselves and their workers higher than necessary wages, we should consider this area as rent—it represents unnecessary payments rather than the misuse of real resources.[32] If, however, potential employees and managers spend time or other resources attempting to secure the rent (say, by enduring periods of unemployment while waiting for one of the overpaid jobs to open up), then the rent may be dissipated and thus converted to deadweight loss.

Note that in the case of an unregulated natural monopoly, one would not expect X-inefficiency to occur unless there were a divergence of information and interests between owners of the firm and its managers. We put this type of problem into the perspective of agency theory in our discussion of government failures in Chapter 8.

In summary, natural monopoly inherently involves the problem of undersupply by the market. The extent of undersupply depends on the particular cost-and-demand conditions facing the monopolist and the extent to which the market can be contested. Natural monopoly may involve additional social surplus losses because the absence of competition permits production at greater than minimum cost to persist.

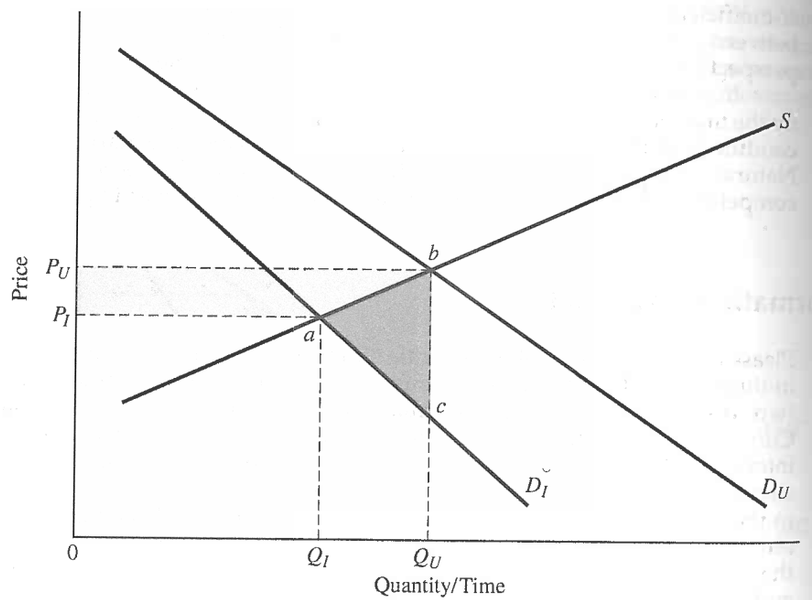
Information Asymmetry

Please note that we do not use the title "information costs" or "imperfect information" in this section. The reason is that information is involved in market failure in at least two distinct ways. First, information itself has public good characteristics. Consumption of information is nonrivalrous—one person's consumption does not interfere with another's; the relevant analytical question is primarily whether exclusion is possible. Thus, in the context of lack of supply of public goods we are interested in the production and consumption of information itself. Second, and the subject of our discussion here, there may be situations in which the amount of information about the characteristics of a good varies in relevant ways across persons. The buyer and the seller in a market transaction, for example, may have different information about the quality of the good being traded. Similarly, there may be differences in the amount of information relating to the attributes of an externality between the generator of the externality and the affected party. Workers, for instance, may not be as well informed about the health risks of industrial chemicals as their employers. Notice that *in the information asymmetry context we are not primarily interested in information as a good, but in differences in the information that relevant parties have about any good.* We thus distinguish between information *as* a good (the public good case) and information *about* a good's attributes as distributed between buyer and seller or between externality generator and affected party (the *information asymmetry* case).

[32]For evidence that unions are successful in capturing some of this rent in the presence of monopoly, see Thomas Karier, "Unions and Monopoly Power," *Review of Economics and Statistics* 67(1) 1985, 34–42.

Inefficiency Due to Information Asymmetry

Figure 5.12 illustrates the potential social surplus loss associated with information asymmetry.[33] D_U represents the quantities of some good that a consumer would purchase at various prices in the absence of perfect information about its quality. It can, therefore, be thought of as the consumer's uninformed demand schedule. D_I represents the consumer's informed demand schedule—the amounts of the good that would be purchased at various prices if the consumer were perfectly informed about its quality. The quantity actually purchased by the uninformed consumer is determined by the intersection of D_U with the supply schedule, S. This amount, Q_U, is greater than Q_I, the amount that the consumer would have purchased if fully informed about the quality of the good. The darkly shaded area *abc* equals the deadweight loss in consumer surplus resulting from the overconsumption. (For each unit purchased beyond Q_I, the consumer pays more than its marginal value as measured



Uninformed consumption: Q_U
Informed consumption: Q_I
Deadweight loss if uninformed: *abc*
Extra producer rent if uninformed: $P_U baP_I$

Figure 5.12 Consumer Surplus Loss from Uninformed Demand

[33]This analysis was introduced by Sam Peltzman, "An Evaluation of Consumer Protection Legislation: The 1962 Drug Amendments," *Journal of Political Economy* 81(5) 1973, 1049–91. For a discussion of the empirical problems in using this approach when some consumers overestimate and others underestimate the quality of some good, see Thomas McGuire, Richard Nelson, and Thomas Spavins, "An Evaluation of Consumer Protection Legislation: The 1962 Drug Amendments: A Comment," *Journal of Political Economy* 83(3) 1975, 655–61.

by the height of the informed demand schedule.) This excess consumption also results in a higher equilibrium price (P_U), which transfers surplus equal to the area $P_U baP_I$ from the consumer to the producer of the good. Figure 5.12 signals the presence of information asymmetry if the producer could have informed the consumer about the true quality of the good at a cost less than the deadweight loss in consumer surplus resulting when the consumer remains uninformed. More generally, *we have market failure due to information asymmetry when the producer does not supply the amount of information that maximizes the difference between the reduction in deadweight loss and the cost of providing the information.*

The same sort of reasoning would apply if the consumer underestimated rather than overestimated the quality of the good. The consumer would suffer a deadweight loss resulting from consuming less than Q_I. The incentives that the producer faces to provide the information, however, can be quite different in the two cases. In the case where the consumer overestimates quality, providing information results in a lower price and, therefore, a smaller transfer of surplus from the consumer to the producer, an apparent disincentive to provide the information. In the case where the consumer underestimates quality, providing information results in a higher price that increases producer surplus; the prospect of this gain may encourage the producer to supply information. As we discuss later, however, this incentive to provide information can be muted if producers are unable to get consumers to distinguish their products from those of competitors.

Diagnosing Information Asymmetry

Our first task in deciding when information asymmetry is likely to lead to market failure is to classify goods into useful categories. Economists have generally divided goods into two categories: *search goods* and *experience goods*.[34] A good is a search good if consumers can determine its characteristics with certainty prior to purchase. For example, a chair in stock in a store is a search good because consumers can judge its quality through inspection prior to purchase. A good is an experience good if consumers can determine its characteristics only after purchase; examples include meals, hairstyling, concerts, legal services, and used automobiles. We add a third category, which we call *post-experience goods*, to distinguish those goods for which it is difficult for consumers to determine quality even after they have begun consumption. For example, people may fail to associate adverse health effects with drugs that they are consuming. Experience goods and post-experience goods differ primarily in terms of how effectively consumers can learn about quality through consumption. After some period of consumption, the quality of the experience goods generally becomes apparent; in contrast, continued consumption does not necessarily reveal to consumers the quality of the post-experience good.[35]

Within these three categories, a number of other factors help determine whether information asymmetry is likely to lead to serious market failure. The effectiveness of any information-gathering strategy, other things equal, generally depends on the *variance in the quality* of units of a good (heterogeneity) and the *frequency* with which consumers make purchases. The potential costs of the information asymmetry to

[34]The distinction between search and experience goods was introduced by Philip Nelson, "Information and Consumer Behavior," *Journal of Political Economy* 78(2) 1970, 311–29.

[35]See Aidan R. Vining and David L. Weimer, "Information Asymmetry Favoring Sellers: A Policy Framework," *Policy Sciences* 21(4) 1988, 281–303.

consumers depend on the extent to which they perceive the *full price* of the good, including imputed costs of harm from use.[36] The *cost of searching* for candidate purchases and the full price determine how expensive and potentially beneficial it is for consumers to gather information.

Search Goods. Searching can be thought of as a sampling process in which consumers incur costs to inspect units of a good. A consumer pays a cost C_S to see a particular combination of price and quality. If the price exceeds the consumer's marginal value for the good, no purchase is made and the consumer either again pays C_S to see another combination of price and quality or stops sampling. If the consumer's marginal valuation exceeds price, then the consumer either makes a purchase or pays C_S again in expectation of finding a more favorable good in terms of the excess of marginal value over price. When C_S is zero, the consumer will find it advantageous to take a large sample and discover the complete distribution of available price and quality combinations so that the pre-search information asymmetry disappears. For larger C_S, however, the consumer will take smaller samples, other things equal, so that information asymmetry may remain. Additionally, because the range in price for a given quality is likely to be positively correlated with price, optimal sample sizes will be smaller the larger the ratio of C_S to expected price.

The more heterogeneous the available combinations of price and quality, the more likely that the consumer will fail to discover a more favorable choice for any given sample size. In contrast, even small samples will eliminate information asymmetry if the price and quality combinations are highly homogeneous. Once consumers realize that nearly identical units are offered at the same price, the optimal sample size falls to one.

Going beyond a static view, the frequency of purchase becomes important in determining whether information asymmetry remains. *If the frequency of purchase is high relative to the rate at which the underlying distribution of combinations of price and quality changes, then consumers accumulate information over time that reduces the magnitude of the information asymmetry.* If the frequency of purchase is low relative to the rate of change in the underlying distribution, then accumulated information will not necessarily lead to reductions in information asymmetry. In either case, however, frequent purchasers may become more experienced searchers so that C_S falls and larger samples become efficient.

Thus, if search costs are small relative to the expected purchase price *or* the distribution of price and quality combinations is fairly homogeneous *or* the frequency of purchase is high relative to the rate of change in the distribution of price and quality combinations, then information asymmetry is unlikely to lead to significant inefficiency. *In the case where search costs are high relative to the expected purchase price, the distribution of price and quality combinations is very heterogeneous, and the frequency of purchase is relatively low, information asymmetry may lead to significant inefficiency.* However, if it is possible for producers to distinguish their products by brand, they have an incentive to undertake informative advertising that reduces search costs for

[36]A formal specification of the concept of full price is provided by Walter Oi, "The Economics of Product Safety," *Bell Journal of Economics and Management Science* 4(1) 1973, 3–28. He considers the situation in which a consumer buys X units of a good at price P per unit. If the probability that any one unit is defective is 1 – q, then, on average, the consumer expects $Z = qX$ good units. If each bad unit inflicts on average damage equal to W, then the expected total cost of purchase is $C = PX + W(X - Z)$, which implies a full price per good unit of $P^* = C/Z = P/q + W(1 - q)/q$.

consumers. When brands are difficult to establish, as in the case of, say, farm produce, retailers may act as agents for consumers by advertising prices and qualities. Because the veracity of such advertising can be readily determined by consumers through inspection of the goods, and because retailers and firms offering brand name products have an incentive to maintain favorable reputations, we expect the advertising generally to convey accurate and useful information. Reputation is likely to be especially important in emerging electronic commerce markets because consumers cannot directly examine goods before purchasing them. For example, one recent study found that sellers' reputation had a statistically significant, albeit small, positive impact on price for an Internet auction good.[37]

In summary, search goods rarely involve information asymmetry that leads to significant and persistent inefficiency. When inefficiency does occur, it takes the form of consumers' forgoing purchases that they would have preferred to ones that they nevertheless found beneficial. From the perspective of public policy, intervention in markets for search goods can rarely be justified on efficiency grounds.

Experience Goods: Primary Markets. Consumers can determine the quality of experience goods with certainty only through consumption. To sample, they must bear the search costs, C_S, and the full price, P^* (the purchase price plus the expected loss of failure or damage collateral with consumption).[38] The full price of consuming a meal at an unfamiliar restaurant, for instance, is the sum of purchase price (determined from the menu) and the expected cost of any adverse health effects (ranging from indigestion to poisoning) from the meal being defective. Of course, even when the expected collateral loss is zero, prior to consumption, the marginal value that the consumer places on the meal is not known with certainty.

In contrast to search goods, where, holding search costs constant, consumers optimally take larger samples for more expensive goods, they optimally take smaller samples for more expensive experience goods. Indeed, for all but the very inexpensive experience goods, we expect sampling (equivalent to the frequency of purchase for experience goods) to be governed primarily by durability. For example, sampling to find desirable restaurants will generally be more frequent than sampling to find good used automobiles.

As is the case with search goods, *the more heterogeneous the quality of an experience good, the greater is the potential for inefficiency due to information asymmetry.* The consumption of an experience good, however, may involve more than simply forgoing a more favorable purchase of the good. Once consumption reveals quality, the consumer may discover that the good provides less marginal value than its price and therefore regret having made the purchase regardless of the availability of alternative units of the good. The realized marginal value may actually be negative if consumption causes harm.

Learning from the consumption of experience goods varies in effectiveness. If the quality of the good is homogeneous and stable, then learning is complete after the first consumption—consumers know how much marginal value they will derive from

[37]Mikhail I. Melnik and James Alm, "Does a Seller's Ecommerce Reputation Matter? Evidence from eBay Auctions," *Journal of Industrial Economics* 50(3) 2002, 337–49.

[38]In our discussion of information asymmetry, we assume that consumers cannot sue producers for damages (in other words, they do not enjoy a property right to safety.) As we discuss in Chapter 10, under framework regulation, tort and contract law often reduce the inefficiency of information asymmetry by deterring parties from withholding relevant information in market transactions.

their next purchase, including the expected loss from product failure or collateral damage. If the quality is heterogeneous or unstable, then learning proceeds more slowly. Unless consumers can segment the good into more homogeneous groupings, say, by brands or reputations of sellers, learning may result only in better ex ante estimates of the mean and variance of their ex post marginal valuations. When consumers can segment the good into stable and homogeneous groupings, repeated sampling helps them discover their most preferred sources of the good in terms of the mean and variance of quality. For example, national motel chains with reputations for standardized management practices may provide travelers with a low-variance alternative to independent motels in unfamiliar locales.

When can informative advertising play an important role in reducing information asymmetry? Generally, informative advertising can be effective when consumers correctly believe that sellers have a stake in maintaining reputations for providing reliable information. A seller who invests heavily in developing a brand name with a favorable reputation is more likely to provide accurate and useful information than an unknown firm selling a new product or an individual owner selling a house or used automobile.

When consumers perceive that sellers do not have a stake in maintaining a good reputation, and the marginal cost of supply rises with quality, then a "lemons" problem may arise: consumers perceive a full price based on average quality so that only sellers of lower than average quality goods can make a profit and survive in the market.[39] In the extreme, producers offer only goods of low quality.[40]

When reliability is an important element of quality, firms may offer warranties that promise to compensate consumers for a portion of replacement costs, collateral damage, or both. The warranties thus provide consumers with insurance against low quality: higher purchase prices include an implicit insurance premium, and the potential loss from product failure is reduced. The quality of a warranty may itself be uncertain, however, if consumers do not know how readily firms honor their promises. Further, firms may find it impractical to offer extensive warranties in situations where it is costly to monitor the care taken by consumers. Nevertheless, warranties are a common device for reducing the consequences of information asymmetry for experience goods.

Experience Goods: The Role of Secondary Markets. Producers and consumers often turn to private third parties to help remedy information asymmetry problems. Certification services, agents, subscription services, and loss control by insurers are the most common market responses that arise.

Certification services "guarantee" minimum quality standards in processes or products. Professional associations, for instance, often set minimum standards of training or experience for their members—the board-certified specialties in medicine are examples. Perhaps closer to most people's experience, the Better Business Bureau

[39]The "lemons" argument originated with George Akerlof, "The Market for Lemons," *Quarterly Journal of Economics* 84(3) 1970, 488–500.

[40]Economists have considered the role of expenditures to establish reputation as a means for high-quality producers to distinguish themselves. See, for instance, Benjamin Klein and Keith B. Leffler, "The Role of Market Forces in Assuring Contractual Performance," *Journal of Political Economy* 89(4) 1981, 615–41; and Paul Milgrom and John Roberts, "Price and Advertising Signals of Product Quality," *Journal of Political Economy* 94(4) 1986, 796–821.

requires members to adhere to a code of fair business practices. Underwriters Laboratories test products against minimum fire safety standards before giving its seal of approval.[41] When such services establish their own credibility, they help producers to distinguish their goods satisfying the minimum standards from goods that do not.

Agents often sell advice about the qualities of expensive, infrequently purchased, heterogeneous goods. These agents combine expertise with learning from being participants in a large number of transactions between different pairs of buyers and sellers. For example, most people do not frequently purchase houses, which are expensive and heterogeneous in quality. Because owners typically enjoy an informational advantage by virtue of their experiences living in their houses, prospective buyers often turn to engineers and architects to help assess structural integrity. Art and antique dealers, jewelers, and general contractors provide similar services.

The problem of excluding nonpaying users' limits the supply of agents for goods that are homogeneous because one consumer can easily pass along information that the agent provides to prospective purchasers of similar units of the good. Inexpensive goods are unlikely to provide an adequate incentive for consumers to pay for the advice of agents. (Note that the full price is the relevant measure: one might very well be willing to pay for a visit to a doctor to get advice about a drug with a low purchase price but a high potential for harm to one's health.) Finally, agents are less likely to be relatively attractive for goods with high frequencies of purchase because consumers can often learn effectively on their own.

Consumers often rely on the experiences of friends and relatives to gather information about the quality of branded products. They may also be willing to pay for published information about such products. But such subscription services, which have public good characteristics, are likely to be undersupplied from the perspective of economic efficiency because nonsubscribers can often free-ride on subscribers by interrogating them and by borrowing their magazines. (The very existence of subscription services such as *Consumer Reports* suggests that a large number of consumers see the marginal costs of mooching or using a library as higher than the subscription price.)

Insurers sometimes provide information to consumers as part of their efforts to limit losses. For example, health maintenance organizations, which provide medical insurance, often publish newsletters that warn members of potentially dangerous and unhealthy goods such as diet products and tanning salons. Casualty insurers may signal warnings through their premiums as well as through direct information. Fire insurance underwriters, for instance, set premiums based on the types of equipment that businesses plan to use; they also often inspect commercial properties to warn proprietors of dangerous equipment and inform them about additions that could reduce their risks of fire.

Experience Goods: Summary. By their very nature, experience goods offer the potential for serious inefficiency caused by information asymmetry. Secondary markets, however, limit the inefficiency for many experience goods by facilitating consumer learning and providing incentives for the revelation of product quality. Nevertheless,

[41]For an overview of private organizations that set quality standards, see Ross E. Cheit, *Setting Safety Standards: Regulation in the Public and Private Sectors* (Berkeley: University of California Press, 1990).

problems are likely to remain in two sets of circumstances: first, when quality is highly heterogeneous, branding is ineffective, and agents are either unavailable or expensive relative to the full price of the good; and second, where the distribution of quality is unstable, so that consumers and agents have difficulty learning effectively. In these fairly limited circumstances, market failure may justify public intervention on efficiency grounds.

Post-Experience Goods. Consumption does not perfectly reveal the true quality of post-experience goods.[42] Quality remains uncertain because the individual consumer has difficulty recognizing the causality between consumption and some of its effects. For example, consumers may not recognize that the fumes of a cleaning product increase their risks of liver cancer because they do not expect the product to have such an effect. Often, the effects of consumption only appear after delay so that consumers do not connect them with the product. Young smokers, for instance, may fully appreciate the addictive power of tobacco only after they are addicted although the manufacturer did know, or should have reasonably known of this effect. Beyond drugs, many medical services appear to be post-experience goods because patients often cannot clearly link the state of their health to the treatments that they have received.

In some cases it is extremely difficult to distinguish between the lack of supply of a public good (the research to find out about the long-term effects of a product) and information asymmetry. An example is DES, a drug that increases the risk of cancer in mature daughters of women who used it during pregnancy. Many other drugs require extended periods before all their effects are manifested. Did the seller know, or could be reasonably expected to know, of the effect at the time of sale? If so, then it is an information asymmetry problem.

In terms of our earlier discussion, the consumer of a post-experience good may not have sufficient knowledge of the product to form reasonable estimates of its full price even after repeated consumption. In other words, whereas experience goods involve risk (known contingencies with known probabilities of occurrence), post-experience goods involve uncertainty (unknown contingencies or unknown probabilities of occurrence). Further, repeated consumption of a post-experience good does not necessarily lead to accurate conversion of uncertainty to risk.

Obviously, the potential for substantial and persistent inefficiency due to information asymmetry is greater for post-experience goods than for experience goods. Generally, *frequent purchases are not effective in eliminating information asymmetry for post-experience goods.* In contrast to the case of experience goods, information asymmetry can persist for an extended period for even homogeneous post-experience goods.

Turning to private responses to the information asymmetry inherent in the primary markets for post-experience goods, we expect secondary markets in general to be less effective than they are for experience goods because of learning problems. Nevertheless, secondary markets can play important roles. For example, consider the

[42]Our category of post-experience goods shares characteristics with Arrow's "trust goods" and captures Darby and Karni's "credence qualities." Kenneth J. Arrow, "Uncertainty and the Welfare Economics of Medical Care," *American Economic Review* 53(5) 1963, 941–73; and Michael R. Darby and Edi Karni, "Free Competition and the Optimal Amount of Fraud," *Journal of Law and Economics* 16(1) 1973, 67–88.

services that are or have been privately offered to provide information about pharmaceuticals. The *Medical Letter on Drugs and Therapeutics*, one of the few medical periodicals that covers its costs through subscriptions rather than advertising, provides independent assessments of all new drugs approved by the FDA. Many hospitals, and some insurers, have formulary committees that review information about drugs for participating physicians. Between 1929 and 1955, the American Medical Association ran a seal of acceptance program that subjected all products advertised in the *Journal of the American Medical Association* to committee review. The National Formulary and the United States Pharmacopoeia were early attempts at standardizing pharmaceutical products. Of course, most of these information sources are directed at physicians and pharmacists who commonly serve as agents for consumers of therapeutic drugs. Although their learning from direct observation may not be effective, they can often determine whether advertising claims for products are based on reasonable scientific evidence.

Reprise of Information Asymmetry

The potential for inefficiency due to information asymmetry between buyers and sellers can be found in many markets. There is some evidence that Web-based, or e-commerce, markets are particularly prone to information asymmetry on some dimensions; for example, consumers cannot touch or feel goods and are, therefore, less likely to be able to assess directly their quality.[43] The potential is rarely great for search goods, often great for experience goods, and usually great for post-experience goods. The extent to which the potential is actually realized, however, depends largely on whether public goods problems hinder the operation of secondary market mechanisms that provide corrective information. Thus, market failure is most likely in situations where information asymmetry in primary markets occurs in combination with public goods problems in secondary markets.

Conclusion

Traditional market failures involve circumstances in which the "invisible hand" fails to produce Pareto efficiency. They thus indicate possibilities for improving the efficiency of market outcomes through collective action. Along with the pursuit of distributional values, which we consider in Chapter 7, they constitute the most commonly advanced rationales for public policy. However, other, less easily accommodated, discrepancies between the competitive framework and the real world also suggest opportunities for advancing efficiency through collective action. These discrepancies, the subject of the next chapter, provide additional rationales for public policy.

[43]For discussions of this issue, see Severin Borenstein and Garth Saloner, "Economics and Electronic Commerce," *Journal of Economic Perspectives* 15(1) 2001, 3–12; and James V. Koch and Richard J. Cebula, "Price, Quality, and Service on the Internet: Sense and Nonsense," *Contemporary Economic Policy* 20(1) 2002, 25–37.

For Discussion

1. Suppose that you and two friends are going to share a house. You have to decide whether each of you will purchase your own groceries and cook your own meals or whether you will purchase groceries and cook meals as a group. What factors might make you more or less likely to choose the group option? Would you be more or less likely to choose it if there were five in the group rather than three?
2. Consider a developing country with a single railroad between the primary port and the most populous city. Under what conditions might the railroad have natural monopoly characteristics?
3. Under what conditions is the market for vaccination against a communicable disease likely to be inefficient?
4. Show the social surplus loss in a situation in which a consumer's uninformed demand schedule lies below her informed demand schedule.

6

Rationales for Public Policy

Other Limitations of the Competitive Framework

Although the four traditional market failures (public goods, externalities, natural monopolies, and information asymmetries) represent violations of the ideal competitive model, we nevertheless were able to illustrate their efficiency consequences with the basic concepts of producer and consumer surplus. The consequences of relaxing other assumptions of the competitive model, however, cannot be so easily analyzed with the standard tools of microeconomics. This in no way reduces their importance. They often serve as rationales, albeit implicitly, for public policies.

We begin by considering two of the most fundamental assumptions of the competitive model: first, participants in markets behave competitively, and second, individual preferences can be taken as fixed, exogenous, and fully rational. We next look at the assumptions that must be made to extend the basic competitive model to uncertain and multiperiod worlds. We then consider the assumption that the economy can costlessly move from one equilibrium to another when circumstances change. Finally, we address briefly the role of macroeconomic policy in managing aggregate economic activity.