Chapter 16

REGULATION OF NATURAL MONOPOLY

PAUL L. JOSKOW*

Department of Economics, Massachusetts Institute of Technology

Contents

* A significant amount of the material in this chapter has been drawn from my lectures on the regulation of natural monopolies in the graduate course that I have taught at MIT for many years. I have had the privilege of teaching this course multiple times with each of my colleagues Nancy Rose, Dick Schmalensee and Jean Tirole. In many cases I can no longer distinguish what came initially from their lectures and what came from mine. To the extent that I have failed to give adequate credit to their contributions, I must apologize and thank them for what they have taught me over the years.

Handbook of Law and Economics, Volume 2
Edited by A. Mitchell Polinsky and Steven Shavell

DOI: 10.1016/S1574-0730(07)02016-6

Abstract

This chapter provides a comprehensive overview of the theoretical and empirical literature on the regulation of natural monopolies. It covers alternative definitions of natural monopoly, public interest regulatory goals, alternative regulatory institutions, price regulation with full information, price regulation with imperfect and asymmetric information, and topics on the measurement of the effects of price and entry regulation in practice. The chapter also discusses the literature on network access and pricing to support the introduction of competition into previously regulated monopoly industries.

Keywords

Natural monopoly, economies of scale, sunk costs, price regulation, public utilities, incentive regulation, performance based regulation, network access pricing

JEL classification: K20, K23, L43, L51, L90

1. Introduction

Textbook discussions of price and entry regulation typically are motivated by the asserted existence of an industry with "natural monopoly" characteristics [e.g. Pindyck and Rubinfeld (2001, p. 50)]. These characteristics make it economical for a single firm to supply services in the relevant market rather than two or more competing. Markets with natural monopoly characteristics are thought to lead to a variety of economic performance problems: excessive prices, production inefficiencies, costly duplication of facilities, poor service quality, and to have potentially undesirable distributional impacts.

Under U.S. antitrust law the possession of monopoly power itself is not illegal. Accordingly, where monopoly "naturally" emerges due to the attributes of the technology for producing certain services, innovation or unique skills, antitrust policy cannot be relied upon to constrain monopoly pricing. Nor are the antitrust laws well suited to responding to inefficiencies resulting from entry of multiple firms in the presence of economies of scale and scope. Accordingly, antitrust policy alone cannot be relied upon to respond to the performance problems that may emerge in markets with natural monopoly characteristics. Administrative regulation of prices, entry, and other aspects of firm behavior have instead been utilized extensively in the U.S. and other countries as policy instruments to deal with real or imagined natural monopoly problems.

American economists began analyzing natural monopolies and the economic performance issues that they may raise over 100 years ago [Lowry (1973), Sharkey (1982), Phillips (1993)] and refinements in the basic concepts of the cost and demand attributes that lead to natural monopoly have continued to evolve over time [Kahn (1970), Schmalensee (1979), Baumol, Panzar, and Willig (1982), Phillips (1993), Laffont and Tirole (1993, 2000), Armstrong, Cowan, and Vickers (1994)]. On the policy side, price and entry regulation supported by natural monopoly arguments began to be introduced in the U.S. in the late 19th century. The scope of price and entry regulation and its institutional infrastructure grew considerably during the first 75 years of the 20th century, covering additional industries, involving new and larger regulatory agencies, and expanding from the state to the federal levels. However, during the 1970s both the natural monopoly rationale for and the consequences of price and entry regulation came under attack from academic research and policy makers [Winston (1993)]. Since then, the scope of price and entry regulation has been scaled back in many regulated industries. Some industries have been completely deregulated. Other regulated industries have been or are being restructured to promote competition in potentially competitive segments and new performance-based regulatory mechanisms are being applied to core network segments of these industries that continue to have natural monopoly characteristics [Winston (1993), Winston and Peltzman (2000), Armstrong and Sappington (2006), Joskow (2006)]. Important segments of the electric power, natural gas distribution, water, and telecommunications industries are generally thought to continue to have natural monopoly characteristics and continue to be subject to price and entry regulation of some form.

Economic analysis of natural monopoly has focused on several questions which, while related, are somewhat different. One question is a normative question: What is the most efficient number of sellers (firms) to supply a particular good or service given firm cost characteristics and market demand characteristics? This question leads to technological or cost-based definitions of natural monopoly. A second and related question is a positive question: What are the firm production or cost characteristics and market demand characteristics that lead some industries "naturally" to evolve to a point where there is a single supplier (a monopoly) or a very small number of suppliers (an oligopoly)? This question leads to behavioral and market equilibrium definitions of natural monopoly which are in turn related to the technological attributes that characterize the cost-based definitions of natural monopoly. A third question is also a normative question: If an industry has "a tendency to monopoly" what are the potential economic performance problems that may result and how do we measure their social costs? This question leads to an evaluation of the losses in economic efficiency and other social costs resulting from an "unregulated" industry with one or a small number of sellers. This question in turn leads to a fourth set of questions: When is government regulation justified in an industry with natural monopoly characteristics and how can regulatory mechanisms best be designed to mitigate the performance problems of concern?

Answering this set of questions necessarily requires both theoretical and empirical examinations of the strengths and weaknesses of alternative regulatory mechanisms. Regulation is itself imperfect and can lead to costly and unanticipated firm responses to the incentives created by regulatory rules and procedures. The costs of regulation may exceed the costs of unregulated naturally monopoly or significantly reduce the net social benefits of regulation. These considerations lead to a very important policy-relevant question. Are imperfect unregulated markets better or worse than imperfectly regulated markets in practice?

Finally, firms with de facto legal monopolies that are subject to price and entry regulation inevitably are eventually challenged by policymakers, customers or potential competitors to allow competing suppliers to enter one or more segments of the lines of business in which they have de facto legal monopolies. Entry may be induced by changes in technology on the costs and demand sides or as a response to price, output and cost distortions created by regulation itself. These considerations lead to a final set of questions. How do changes in economic conditions or the performance of the institution of regulated monopoly lead to public and private interests in replacing regulated monopoly with competition? How can policymakers best go about evaluating the desirability of introducing competition into these industries and, if competition appears to be desirable, fashioning transition mechanisms to allow it to evolve efficiently?

Scholarly law and economics research focused on answering these positive and normative questions has involved extensive theoretical, empirical, and institutional analysis. Progress has been made as well through complementary research in law, political sciences, history, organizational behavior and corporate finance. This chapter adopts a similarly comprehensive perspective of the research on the natural monopoly problem relevant to a law and economics handbook by including theoretical, empirical,

policy and institutional research and identifying linkages with these other disciplines. Indeed, research on economic regulation has flourished because of cooperative research efforts involving scholars in several different fields. Nevertheless, the Chapter's primary perspective is through the lense of economic analysis and emphasizes the economic efficiency rationales for and economic efficiency consequences of government regulation of prices and entry of firms producing services with natural monopoly characteristics. In addition, several industries have been subject to price and entry regulation which clearly do not have natural monopoly characteristics (e.g. trucking, natural gas and petroleum production, airlines, agricultural commodities). These multi-firm regulated industries have been studied extensively and in many cases have now been deregulated [Joskow and Noll (1981), Joskow and Rose (1989)]. This chapter will not cover regulation of multi-firm industries where natural monopoly is an implausible rationale for regulation.

The chapter proceeds in the following way. The first substantive section discusses alternative definitions of natural monopoly and the attributes of technologies, demand and market behavior that are thought to lead to natural monopolies from either a normative or a positive (behavioral) perspective. The section that follows it examines the rationales for introducing price and entry regulation in sectors that are thought to have natural monopoly characteristics. This section enumerates the economic performance problems that may result from natural monopoly, focusing on economic efficiency considerations while identifying equity, distributional and political economy factors that have also played an important role in the evolution of regulatory policy. This discussion leads to a set of normative goals that are often defined for regulators that reflect these performance problems. Section 4 provides a brief discussion of the historical evolution of and legal foundations for price and entry regulation, emphasizing developments in the U.S. Section 5 discusses alternative institutional frameworks for regulating legal monopolies, including direct legislative regulation, franchise contracts, and regulation by independent regulatory commissions.

The chapter then turns to a discussion of optimal regulatory mechanisms given different assumptions about the information available to the regulator and the regulated firm and various economic and legal constraints. Section 6 discusses optimal price regulation of a monopoly with subadditive costs in a world where the regulator is perfectly informed about the regulated firm's costs and has the same information about the attributes of demand faced by the regulated firm as does the firm. This section includes a discussion of Ramsey-Boiteux pricing, two-part tariffs, more general models of nonlinear pricing, and peak load pricing. The section that follows it begins a discussion of regulatory mechanisms in a world where the regulator has limited or imperfect and asymmetric information about the attributes of the regulated firm's cost opportunities, the attributes of consumer demand for its services and the managerial effort exerted by its managers. It discusses how traditional cost-of-service regulation evolved in an effort to reduce the regulator's information disadvantage and the early analytical models that sought to understand the efficiency implications of cost of service or rate of return regulation. This discussion sets the stage for a review of the more recent theoretical literature on incentive or performance based regulation where the regulator has imperfect

and asymmetric information about firm's cost opportunities, demand, and managerial effort attributes and the basic practical lessons that can be learned from it. Section 9 turns to recent empirical research that seeks to measure the effects of price and entry regulation of legal monopolies using a variety of performance indicia. The section focuses on post-1990 research on the effects of incentive regulation in practice. Earlier empirical research is discussed in Joskow and Rose (1989).

Individual vertical segments or lines of business of many industries that had been regulated as vertically integrated monopolies for many years have been opened up to competition in recent years (e.g. intercity telecommunications, electricity generation, natural gas production) as remaining "network infrastructure" segments remain regulated and provide a platform for competition in the potentially competitive segments. The introduction and success of competition in one or more of these vertical segments often involves providing access to network facilities that continue to be controlled by the incumbent and subject to price regulation. Accordingly, introducing competition in these segments requires regulators to define the terms and conditions of access to these "essential" network facilities and ensure that they are implemented. Section 10 discusses theoretical research on competitive entry and network access pricing. A brief set of conclusions completes the chapter.

2. Definitions of natural monopoly

2.1. Technological definitions of natural monopoly

I have not been able to determine definitively when the term "natural monopoly" was first used. Sharkey (1982, pp. 12–20) provides an excellent overview of the intellectual history of economic analysis of natural monopolies and I draw on it and the references he sites here and elsewhere in this chapter. He concludes [Sharkey (1982, p. 14)] that John Stuart Mill was the first to speak of natural monopolies in 1848. In his *Principles of Economics*, Alfred Marshall (1890) discusses the role of "increasing returns" in fostering monopoly and oligopoly, though he appears to be skeptical that pure monopolies can endure for very long or profitably charge prices that are significantly above competitive levels without attracting competitive entry [Marshall (1890, pp. 238–239, 329, 380)]. Posner (1969, p. 548) writes that natural monopoly "does not refer to the actual number of sellers in a market but to the relationship between demand and the technology of supply." Carlton and Perloff (2004, p. 104) write that "When total production costs would rise if two or more firms produced instead of one, the single firm in a market is called a "natural monopoly."

These are simple expositions of the technological definition of natural monopoly: a firm producing a single homogeneous product is a natural monopoly when it is less costly to produce any level of output of this product within a single firm than with two or more firms. In addition, this "cost dominance" relationship must hold over the full range of market demand for this product $Q = D(p)$.

Consider a market for a homogeneous product where each of k firms produces output q^i and total output is given by $Q = \sum_k q^i$. Each firm has an identical cost function $C(q^i)$. According to the technological or cost-based definition of natural monopoly, a natural monopoly will exist when:

$$C(Q) < C(q^1) + C(q^2) + \cdots + C(q^k)$$

since it is less costly to supply output Q with a single firm rather than splitting production up between two or more competing firms. Firm cost functions that have this attribute are said to be *subadditive* at output level Q (Sharkey, 1982, p. 2). When firm cost functions have this attribute for all values of Q (or all values consistent with supplying all of the demand for the product $Q = D(p)$) then the cost function is said to be *globally subadditive*. As a result, according to the technological definition of natural monopoly, a necessary condition for a natural monopoly to exist for output Q of some good is that the cost of producing that good is subadditive at Q.

Assume that firm i's cost function is defined as:[1]

$$C^i = F + cq^i$$

then the firm's average cost of production

$$\mathrm{AC}^i = F/q^i + c$$

declines continuously as its output expands. When a firm's average cost of production declines as its output expands its production technology is characterized by *economies of scale*. A cost function for a single-product firm characterized by declining average total cost over the relevant range of industry output from 0 to $q^i = Q$ is subadditive over this output range. Accordingly, in the single product context, economies of scale over the relevant range of q is a sufficient condition to meet the technological definition of natural monopoly. Figure 1 depicts the cost function for a firm with economies of scale that extend well beyond the total market demand (Q) depicted by the inverse demand function $P = D(Q)$. We note as well that when there are economies of scale up to firm out level q it will also be the case that average cost will be greater than marginal cost over this range of output ($F/q^i + c > c$ in the simple example above).[2]

In the single product case, economies of scale up to $q^i = Q$ is a *sufficient* but not a *necessary* condition for subadditivity over this range or, by the technological definition, for natural monopoly. However, it may still be less costly for output to be produced in a single firm rather than multiple firms even if the output of a single firm has expanded

[1] It should be understood that cost functions utilized here are technically $C = C(q, \mathbf{w})$ where $\mathbf{w}$ is a vector of input prices that we are holding constant at this point. They also reflect cost-minimization by the firm in the sense that the marginal rate of transformation of one input into another is equal to the associated input price ratio.

[2] Some definitions of natural monopoly assert that the relevant characteristic is declining marginal cost. This is wrong.

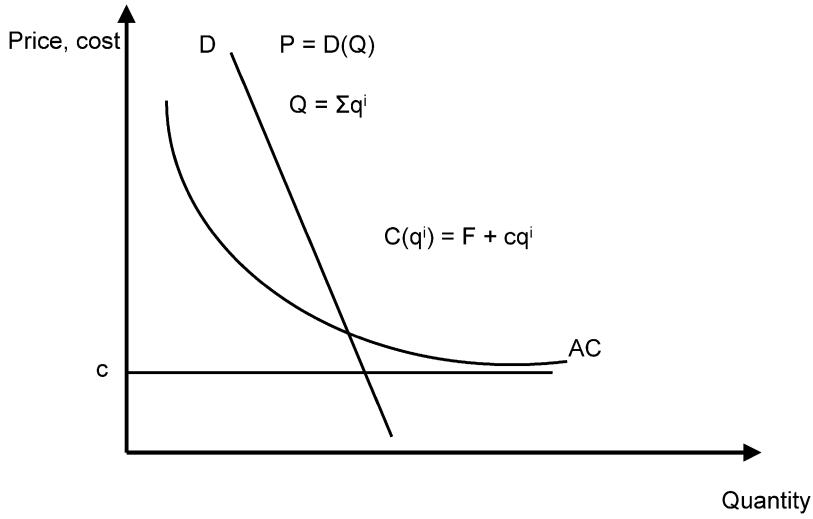


Figure 1. Economies of scale.

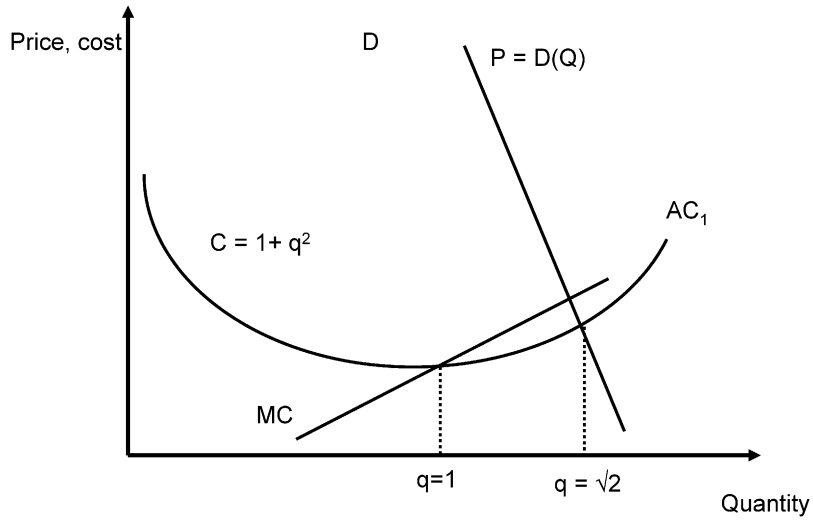


Figure 2. Subadditivity and diseconomies of scale.

beyond the point where there are economies of scale. Consider the total cost function for a firm $C = 1 + q^2$ and the associated average cost function $\text{AC} = q + 1/q$ depicted as AC_1 in Figure 2. There is a range of output where there are economies of scale ($q < 1$). The cost function then flattens out ($q = 1$) and then enters a range of decreasing returns

to scale ($q > 1$). However, this cost function is still subadditive for some values of $q > 1$, despite the fact that for $q > 1$ there is decreasing returns to scale. This is the case because the market demand $P = D(Q)$ is not large enough to support efficient production by two firms for some levels of industry output $Q > 1$.

Assume that firm 1 produces $q^1 = 1$ to produce at minimum efficient scale. Consider a second firm 2 with the same costs that could also produce at minimum efficient scale $q^2 = 1$. If both firms produced at minimum efficient scale total output would be 2 and total cost would be 4. If a single firm produced output $q = 2$, total cost would be 5, so it is more efficient to produce total industry output $Q = 2$ with two firms rather than one. However, it is apparent that for total output levels between $Q = 1$ and $Q = \sqrt{2}$ it is less costly to allow the first firm to operate in a range of decreasing returns to scale than it is to supply with two firms, both producing at greater than minimum efficient scale. Similarly for $q > \sqrt{2}$, it is less costly to supply with two firms rather than one and the cost function is not subadditive in this range.

Accordingly, the set of firm cost functions that are subadditive encompasses a wider range of cost functions than those that exhibit economies of scale over the entire (relevant) range of potential industry output. Specifically, in the single product case, the firm's cost functions must exhibit economies of scale over some range of output but it will still be subadditive in many cases beyond the point where economies of scale are exhausted and until industry output is large enough to make it economical to add a second firm.

There are some implicit assumptions regarding the firm's cost function $C(q)$ that should be noted here. First, it is a long run "economic cost function" in the sense that it reflects the assumption that the firm produces any particular output efficiently, given the underlying production function and input prices, and that inputs are fully adjusted to prevailing input prices and the quantity produced. That is, there is no "X-inefficiency" reflected in the firm's costs. Capital related costs in turn reflect the firm's opportunity cost of capital (r), economic depreciation (d), and the value of the capital invested in productive assets (K) measured at the current competitive market value of the associated assets. That is, the firm's total costs of production include the current period rental cost of capital $V = (r + d)K$. Accordingly, capital costs are not treated explicitly as being sunk costs for the technological definition of natural monopoly. These implicit assumptions have important implications for a variety of issues associated with behavioral definitions of natural monopoly, the measurement of the social costs of unregulated natural monopolies, the social costs of regulation, and the design of effective regulatory mechanisms. I will turn to these issues presently.

The technological definition of natural monopoly can be generalized to take account of multiproduct firms. For this purpose, multiproduct firms are firms that have technologies that make it more economical to produce two or more products within the same firm than in two or more firms. Production technologies with this attribute are characterized by *economies of scope*. Consider two products q_1 and q_2 that can be produced by a firm with a cost function $C(q_1, q_2)$. Define $\mathbf{q}^i$ as a vector of the two products $\mathbf{q}^i = (q_1^i, q_2^i)$. There are N vectors of the two products with the attribute that $\sum q_1^i = q_1$

and $\sum q_2^i = q_2$. Then the cost function $C(q_1, q_2)$ is subadditive if:

$$C\left(\sum q_1^i, \sum q_2^i\right) = C\left(\sum \mathbf{q}^i\right) < \sum C(\mathbf{q}^i)$$

for all N vectors of the products. This definition can be generalized to any number of products.

What attributes of a production technology/cost function will lead to multiproduct subadditivity? The technology must be characterized by some form of *economies of scope* and some form of *multiproduct economies of scale*.

By *economies of scope* we mean that it is more economical to produce the two products in one firm rather than multiple firms:

$$C(q_1, q_2) < C(q_1, 0) + C(0, q_2)$$

There are several concepts of *multiproduct economies of scale* depending upon how one slices the multiproduct cost function:

a. *Declining average incremental cost* for a specific product
b. *Declining ray average cost* for varying quantities of a set of multiple products that are bundled in fixed proportion

Define the *incremental cost* of producing product q_1 holding q_2 constant as

$$\text{IC}(q_1|q_2) = c(q_1, q_2) - c(0, q_2)$$

and define the average incremental cost of producing q_1 as

$$\text{AIC}(q_1|q_2) = \left[c(q_1, q_2) - c(0, q_2)\right]/q_1$$

If the AIC declines as the output of q_1 increases (holding q_2 constant) then we have *declining average incremental* cost of q_1. This is a measure of *single product economies of scale* in a multiproduct context. We can perform the same exercise for changes in q_2 holding q_1 constant to determine whether there are declining average incremental costs for q_2 and, in this way, determine whether the cost function is characterized by *declining average incremental cost* for each product.

We can think of fixing the *proportion* of the multiple products that are produced at some level (e.g. $q_1/q_2 = k$) in the two-product case) and then ask what happens to costs as we increase the quantity of both outputs produced holding their relative output proportions constant. Does the average cost of the bundle decline as the size of the bundle (holding the output proportions constant) increases?

Let λ be a number greater than one. If the total costs of producing this "bundle" of output increase less than proportionately with λ then there are multiproduct economies of scale along a ray defined by the product proportions k. This is called *declining ray average costs* for $q_1/q_2 = k$.

$$c(q_1, q_2|q_1/q_2 = k) > c(\lambda q_1, \lambda q_2|q_1/q_2 = k)/\lambda$$

By choosing different proportions of the products produced (alternative values for k) by the firm we can trace out the cost functions along different rays in q_1, q_2 space and

determine whether there are economies of scale or declining ray average costs along each ray. Then there are *multiproduct economies of scale* in the sense of declining ray average cost for any combination of q_1 and q_2 when:

$$C(\lambda q_1, \lambda q_2) < \lambda C(q_1, q_2)$$

For example, consider the cost function (Sharkey, 1982, p. 5)

$$C = q_1 + q_2 + (q_1 q_2)^{1/3}$$

This cost function exhibits multiproduct economies of scale since

$$\lambda C(q_1, q_2) = \lambda q_1 + \lambda q_2 + \lambda (q_1 q_2)^{1/3}$$
$$C(\lambda q_1, \lambda q_2) = \lambda q_1 + \lambda q_2 + \lambda^{2/3} (q_1 q_2)^{1/3}$$

and thus

$$C(\lambda q_1, \lambda q_2) < \lambda C(q_1, q_2)$$

However, this cost function exhibits *diseconomies of scope* rather than economies of scope since:

$$C(q_1, 0) = q_1$$
$$C(0, q_2) = q_2$$
$$C(q_1, 0) + C(0, q_2) = q_1 + q_2 < q_1 + q_2 + (q_1 q_2)^{1/3} = c(q_1, q_2)$$

As a result, this multiproduct cost function is not subadditive despite the fact that it exhibits declining ray average cost. It would be less costly to produce the two products in separate firms.

Subadditivity of the cost function, or natural monopoly, in the multiproduct context requires both a form of multiproduct cost complementarity (e.g. economies of scope[3]) and a form of multiproduct economies of scale over at least some range of the output of the products. For example, the multiproduct cost function [discussed by Sharkey (1982, p. 7)

$$C(q_1, q_2) = (q_1)^{1/4} + (q_2)^{1/4} - (q_1 q_2)^{1/4}$$

exhibits economies of scope. It also exhibits economies of scale in terms of both declining average incremental cost and declining ray average cost at every level of output of the two products. It is obvious that costs are lower when the products are produced together rather than separately by virtue of the term $(-(q_1 q_2)^{1/4})$ in the cost function. There are also declining ray average cost and declining average incremental cost for each product. This is the case because the cost of producing a particular combination of the two outputs increases less than proportionately with increases in the scale of the

[3] Or one of a number of other measures of cost complementarity.

bundle of two products produced by virtue of the power 1/4 in the cost function. Similarly for the average incremental cost of q_1 and q_2 individually. It can be shown that this cost function is subadditive at every output level or *globally subadditive*.

The necessary and sufficient conditions for global subadditivity of a multiproduct cost function are complex and it is not particularly useful to go into those details here. Interested readers should refer to Sharkey (1982) and to Baumol, Panzar, and Willig (1982). As already discussed, economies of scope is a necessary condition for a multiproduct cost function to be subadditive. One set of sufficient conditions for subadditivity of a multiproduct cost function is that it exhibit both economies of scope and declining average incremental cost for all products. An alternative set of sufficient conditions is that the cost function exhibit both declining average incremental cost for all products plus an alternative measure of multiproduct cost complementarity called *trans-ray convexity*. Trans-ray convexity requires that multiproduct economies outweigh any single product diseconomies of scale. For example, it may be that there are single product economies of scale for product 1, diseconomies of scale for product 2, but large multiproduct economies. Then it could be less costly to produce q_1 and q_2 together despite the diseconomies of scale in producing q_2 to take advantage of the multiproduct economies available from joint production. A third alternative sufficient condition is that the cost function exhibit *cost complementarity*, defined as the property that increased production of any output reduces (does not increase) marginal costs of all other outputs. As in the case of single product cost functions, the necessary conditions regarding scale economies are less strict and allow for output to expand into a range of diseconomies of scale or diseconomies of scope since it may be less costly to produce at a point where there are diseconomies than it is to incur the costs of suboptimal production from a second firm.

2.2. *Behavioral and market equilibrium considerations*

The previous section discussed the attributes of a firm's cost function that would make it most efficient from a cost of production perspective (assuming costs are minimized given technology and input prices as discussed earlier) to concentrate production in a single firm rather than in multiple firms. However, the intellectual evolution of the natural monopoly concept and public policy responses to it focused much more on the consequences for unregulated market outcomes of production technologies having such "natural monopoly" attributes. Moreover, historical discussions of the natural monopoly problem focus on more than economies of scale and related multiproduct cost complementarity concepts as potential sources of market distortions. Sharkey (1982, pp. 12–20) discusses this aspect of the intellectual history of economic analysis of natural monopoly as well. For example, in addition to economies of scale he notes that Thomas Farrer (1902) [referenced by Sharkey (1982, p. 15)] associated natural monopoly with supply and demand characteristics that included (a) the product or service supplied must be essential, (b) the products must be non-storable, (c) the supplier must have a favorable production location. In addition, Richard Ely (1937) [referenced

by Sharkey (1982, p. 15)] added the criteria that (a) the proportion of fixed to variable costs must be high and (b) the products produced from competing firms must be close substitutes. Bonbright (1961, pp. 11–17) suggested that economies of scale was a sufficient but not a necessary condition for natural monopoly and Posner (1969, p. 548) observed that "network effects" could lead to subadditive costs even if the cost per customer increased as the number of customers connected to the network increased; as more subscribers are connected to a telephone network, the average cost per subscriber may rise, but it may still be less costly for a single firm to supply the network service. Kaysen and Turner (1959, pp. 191, 195–196) note that economies of scale is a relative concept that depends on the proper definition of the relevant product and geographic markets and also argue that "ruinous competition" leading to monopoly may occur when the ratio of fixed to variable costs is high and identify what we would now call "sunk costs" as playing an important role leading to monopoly outcomes. Kahn (1970, pp. 119, 173) refers to both economies of scale and the presence of sunk or fixed costs that are a large fraction of total costs as attributes leading to destructive competition that will in turn lead a single firm or a very small number of firms in the market in the long run. He also recognizes the potential social costs of "duplicated facilities" when there are economies of scale or related cost-side economic attributes that lead single firm production to be less costly than multiple firm production.

These expanded definitions of the attributes of natural monopoly appear to me to confuse a set of different but related questions. In particular, they go beyond the normative concept of natural monopoly as reflecting technological and associated cost attributes that imply that a single firm can produce at lower cost than multiple firms, to examine the factors that "naturally" lead a market to evolve to a point where there is a single supplier (or not). That is, they include in their definition of natural monopoly the answer to the positive or behavioral question of what cost and demand attributes lead industries to evolve so that only a single firm survives in the long run? To some extent, some of these definitions also begin to raise normative questions about the consequences of the dynamics of the competitive process for costs, prices, and other aspects of social welfare in industries with natural monopoly characteristics. For example, Kaysen and Turner (1959, p. 191) associated natural monopoly that "leaves the field to one firm ... competition here is self-destructive." They go on to assert that "The major prerequisites for competition to be destructive are fixed or sunk costs that bulk large as a percentage of total costs" [Kaysen and Turner (1959, p. 173)]. Kahn (1970) observes that sunk costs must be combined with significant economies of scale for monopoly to "naturally" emerge in the market. So, the historical evolution of the natural monopoly doctrine reflects both a normative interest in identifying situations in which a single firm is necessary to achieve all economies of scale and multiproduct cost complementarities as well as a positive interest in identifying the attributes of costs and demand that lead to market conditions that are "unsuitable for competition" to prevail and the associated normative performance implications for prices, costs and other attributes of social welfare.

Absent regulatory constraints on pricing and entry, the presence of subadditive costs per se do not necessarily lead to the conclusion that a single firm—a monopoly—will "naturally" emerge in equilibrium. And if a monopoly "naturally" does emerge in equilibrium, a variety of alternative pricing patterns may result depending on cost, demand, and behavioral attributes that affect opportunities for price discrimination, competitive entry and the effects of potential entry on incumbent behavior. After all, many models of imperfect competition with two or more firms are consistent with the assumption that the competing firms have cost functions that are characterized by economies of scale over at least some range of output. Nor, as we shall see presently, if a single firm emerges in equilibrium is it *necessarily* the case that it will charge prices that yield revenues that exceed a breakeven level. On the other, hand, if a single firm (or a small number of firms) emerges in equilibrium it may have market power and charge prices that yield revenues that exceed the breakeven level for at least some period of time, leading to lower output and higher unit costs than is either first-best or second-best efficient (i.e. given a break-even constraint).

In order to draw positive conclusions about the consequences of subadditive costs for the attributes of short run and long run firm and market behavior and performance we must make additional assumptions about other attributes of a firm's costs, the nature of competitive interactions between firms *in* the market and interactions between firms in the market with potential entrants *into* the market when the firm's long run production costs are subadditive. Moreover, if more than one firm survives in equilibrium—e.g. a duopoly—the equilibrium prices, quantities and costs may be less desirable from an economic performance perspective than what is theoretically *feasible* given the presence of subadditive costs and other constraints (e.g. breakeven constraints). This latter kind of result is the foundation for arguments for introducing price and entry regulation in industries with natural monopoly characteristics despite the fact that multiple firms may survive in equilibrium and compete, but compete imperfectly.

2.3. Sunk costs

The most important cost attribute that is not reflected explicitly in the traditional technological definitions of natural monopoly that turn on the presence of subadditive firm production costs is the existence and importance of sunk costs. Sunk cost considerations also provide the linkage between subadditivity, behavioral definitions of natural monopoly, and the economic performance problems that are thought to arise from unregulated natural monopolies. Sunk costs are associated with investments made in long-lived physical or human assets whose value in alternative uses (i.e. to produce different products) or at different locations (when transportation costs are high) is lower than in its intended use. At the extreme, an investment might be worthless in an alternative use. Sunk costs are a "short run" cost concept in the sense that the associated assets eventually are valueless in their intended use and are retired. However, because the assets are long-lived, the short run may be quite long from an economic perspective. Sunk costs are not directly captured in long run neoclassical cost functions since

these cost functions reflect the assumption that capital assets can be rented on a period by period basis and input proportions are fully adjusted to prevailing input prices and output levels. Accordingly, sunk costs have not been considered directly in technological definitions of natural monopoly that turn only on cost subadditivity. Yet, sunk costs are quite important both theoretically and empirically for obtaining a comprehensive understanding of the natural monopoly problem as it has emerged in practice. Sunk cost considerations are important both to explain why some industries "naturally" evolve to a point where one or a very small number of firms survive and to measure the social welfare consequences of the market structures and associated, price, cost and quality attributes of these markets in the absence of price and entry regulation (Sutton, 1991). As discussed below, sunk cost considerations are also important for establishing regulated prices for incumbents when their industries are opened up to competition [Hausman (1997), Pindyck (2004)].

Most of the industries that have been regulated based on natural monopoly arguments—railroads, electric power, telephone, gas pipelines, water networks, cable television networks, etc.—have the attribute that a large fraction of their total costs are sunk capital costs. Moreover, it has been argued that a meaningful economic definition of economies of scale requires that there be at least some sunk costs and, for these purposes, thinking about there being fixed costs without there also being sunk costs is not particularly useful (Weitzman, 1983). Indeed, Weitzman argues that sunk costs introduce a time dimension into the cost commitment and recovery process that is essential to obtaining a useful concept of economies of scale. I will return to this issue presently.

2.4. Contestible markets: subadditivity without sunk costs

In order to get a better feeling for the importance of sunk cost and the behavioral attributes of firms in the market and potential entrants into the market, it is useful to focus first on the model of *contestable markets* developed by Baumol, Panzar, and Willig (1982) which assumes that costs are subadditive but generally ignores sunk costs. The examples that follow will focus on a single product case, but the extension to multiple products is straightforward, at least conceptually. Consider the single product situation in which there are n identical firms (where n is large) with identical cost functions $C(q^i) = F + cq^i$. This cost function is assumed to exhibit economies of scale over the entire range of q and thus is subadditive. One of the n firms (the incumbent) is in the market and the remaining $(n-1)$ firms are potential entrants. The declining average cost curve for the firm in the market is depicted in Figure 3 along with the inverse *market* demand for the product $p = D(q)$ (where the market demand is $Q = \sum q_i = D(p)$). F is assumed initially to be a fixed cost but not a sunk cost. It is not a sunk cost in the sense that firms can enter or exit the market freely without facing the risk of losing any of these fixed costs up to the point in time that the firm actually produces output q^i and incurs operating costs cq^i. If prices are not high enough to cover both a firm's operating cost cq^i and its associated fixed cost F, the firm will either not enter the market or will exit the market before committing to produce and avoiding incurring the associ-

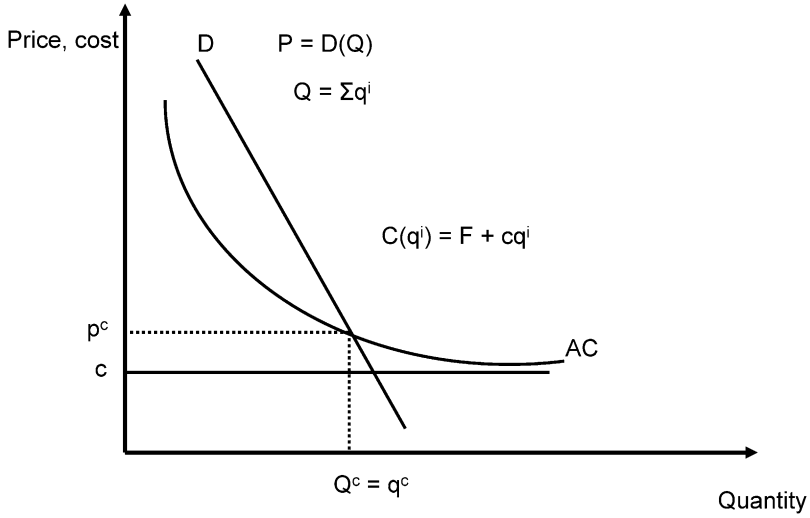


Figure 3. Economies of scale and break-even price.

ated costs. Thus, assuming that fixed costs are not sunk costs is equivalent to assuming that there is hyper-free entry and exit into and out of this market—there are no fixed commitment costs prior to actual production and the fixed costs of production can be avoided by a firm that has "entered" the market by simply not producing any output and effectively exiting the market without incurring any entry or exit costs.[4]

We are looking for an equilibrium where it is (a) *profitable* for one or more firms to enter (or remain in) the market and produce output ($pq^i \geq C(q^i)$), (b) *feasible* in the sense that supply and demand are in balance ($\sum q^i = Q = D(p)$), and (c) *sustainable* in the sense that no entrant can make a profit given the price charged by the incumbent(s)—there does not exist a price $p^a < p$ and an output $Q^a \leq D(p^a)$ such that $p^a q^a \geq C(q^a)$.

Figure 3 depicts an equilibrium that satisfies these conditions. At price p^c and output Q^c ($Q^c = q^c$) the incumbent firm exactly covers its costs and earns zero economic profit since $p^c = F/q^c + c = \mathrm{AC}_c$. It is not profitable for a second firm to enter with a price lower than p^c since it could not break even at any output level at a price less that p^c. The incumbent cannot charge a price higher than p^c (that is, p^c is not sustainable) because if the incumbent committed to a higher price one of the potential entrants could profitably offer a lower price, enter the market and take all of the incumbent's sales away. Moreover, with the incumbent committing to $p > p^c$, competition among potential entrants would drive the price down to p^c and it would be profitable for only

[4] Weitzman (1983) argues that there are no economies of scale in any meaningful sense in this case. Also see Tirole (1988, p. 307). This issue is discussed presently.

one of them to supply in equilibrium due to economies of scale. So, under these conditions the industry equilibrium is characterized by a single firm (a "natural monopoly"). However, the price p^c is the lowest uniform per unit (linear) price consistent with a firm breakeven (zero profit) constraint; the equilibrium price is not the classical textbook monopoly price but the lowest uniform per unit price that allows the single firm producing output to just cover its total costs of production. This price and output configuration is both feasible and sustainable. Thus, the threat of entry effectively forces the single incumbent supplier to charge the lowest uniform (linear) per unit price consistent with a breakeven constraint. As we shall see below, this equilibrium price which is equal to average total cost at the quantity that clears the market is the second-best efficient uniform (Ramsey-Boiteux) price when the firm is subject to a break-even constraint. Obviously, as I shall discuss in more detail below, it is not first best since the equilibrium price is greater than marginal cost (c).

These are remarkable results. They suggest that even with significant increasing returns we "naturally" get to a competitive equilibrium characterized by both a single firm exploiting the cost savings associated with global subadditivity and the lowest price that just allows a single firm exploiting all economies of scale to break even. This is as close to efficient uniform per unit (linear) pricing as we can expect in a market with private firms that are subject to a break-even constraint and have cost functions characterized by economies of scale. The classical textbook problem of monopoly pricing by an incumbent monopoly does not emerge here in equilibrium. In this case potential competition is extremely effective at constraining the ability of the incumbent to exercise market power when it sets prices, with no regulatory intervention at all. If this situation accurately reflected the attributes of the industries that are generally thought of as having "natural monopoly" characteristics then they would not appear to be particularly interesting targets for regulatory intervention (see the next section) since a fully informed regulator relying on uniform per unit prices could do no better than this.

Note, that even in this peculiar setting, an equilibrium with these attributes may not be *sustainable*. Consider the average cost function depicted in Figure 4 that has increasing returns up to point q^o and then enters a range of decreasing returns (perhaps due to managerial inefficiencies as the firm gets very large). The market demand curve crosses the average cost curve at the output level q^a and the average cost at this output level is equal to AC_a. In this case, the price that allows the single firm supplying the entire market to break even and that balances supply and demand is $p_a = AC_a$. However, this price is not sustainable against free entry. An entrant could, for example, profitably enter the market by offering to supply q^o at a price p_o equal to $AC_o + \varepsilon$. In this case, the entrant would have to ration demand to limit its output to q_o. The incumbent could continue to supply to meet the demand that has not been served by the new entrant, but would incur very high average costs to do so and would have to charge higher prices to break even. If we assume that the entrant supplied the consumers with the highest willingness to pay, there would not be any consumers willing to pay a price high enough for the incumbent to cover its average costs. Thus, the zero profit "natural monopoly" equilibrium is unstable.

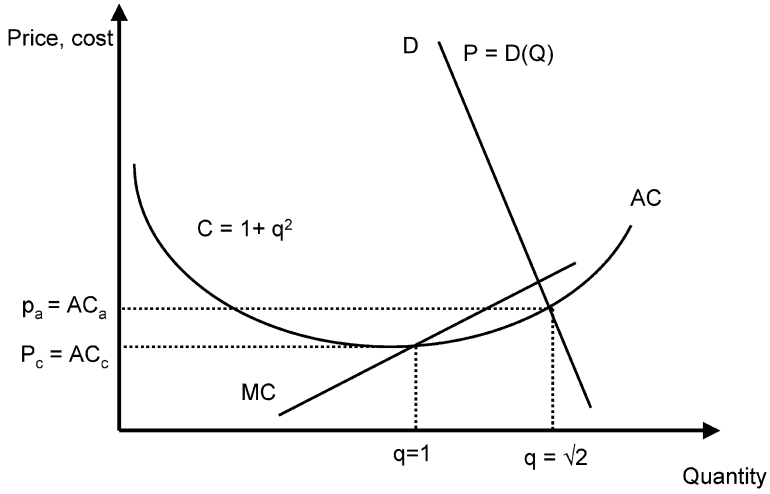


Figure 4. Subadditivity and diseconomies of scale.

In a multiproduct context, perfectly contestable markets (no sunk costs, free entry) have a symmetrical set of attributes. Following, Baumol, Panzar, and Willig (1982), if a sustainable allocation exists, it has the following attributes: (a) there is a single firm to take advantage of cost subadditivity, (b) the firm earns zero profits, (c), the revenues that a firm earns from any subset of products is greater than or equal to the incremental cost of producing that subset of products—there is no "cross-subsidization" in the sense that the prices charged for any product or set of products covers the *incremental* costs incurred to produce them, (d) the price of each product exceeds its (single product) marginal cost given the output of the other products, (e) under certain conditions the firm will voluntarily charge the second-best linear (Ramsey-Boiteux) prices [Baumol, Bailey, and Willig (1977)]. As in the case of a single product firm, the existence of a subadditive multiproduct cost function does not guarantee that a sustainable single-firm zero profit (break-even) configuration exists.

It seems to me that the primary point that emerges from the lengthy literature on contestable markets is that one cannot conclude that there are necessarily "monopoly problems" from the observation that there is one or a very small number of firms producing in a market. Prices may still be competitive in the second best sense ($P = \text{AC}$) in the presence of increasing returns because entry is so easy that it constrains the incumbent's prices. A monopoly naturally emerges, but it may have no or small social costs compared to feasible alternative allocations.

2.5. Sunk costs and barriers to entry

As I have already noted, the assumption that there are fixed costs but no sunk costs does not make a lot of economic sense [Weitzman (1983), Tirole (1988, p. 307)]. Sunk costs

introduce a time dimension into the analysis since sunk costs convey a stream of potential benefits over some period of time and once the associated cost commitments are made they cannot be shifted to alternative uses without reducing their value from that in the intended use. Sunk costs are what make the distinction between incumbents and potential entrants meaningful. Absent sunk costs there is no real difference between firms in the market and firms that are potentially in the market since entry and exit are costless. Sunk costs also create potential opportunities for strategic behavior by the incumbent designed both to sustain prices about the break-even level while simultaneously discouraging entry. If the fixed costs are fully avoidable up to the point that production actually takes place, a firm incurs no opportunity cost merely by entering the market. Whether a firm is "in" the market or "out" of the market is in some sense irrelevant in this case since there is no time dimension to the fixed costs. Firms are only "in" when they start to produce and can avoid incurring any fixed costs if they don't. From an entry and exit perspective, all costs are effectively variable over even the shortest time period relevant for determining prices and output.

An alternative approach that retains the notion that fixed costs are also at least partially sunk involves specifying a price competition game in which fixed cost (capacity) commitments can be adjusted more quickly than can the prices set by the firm and the associated quantities it commits to see [Tirole (1988, pp. 310–311)]. The fixed costs are sunk, but they are sunk for a shorter period of time than it takes to adjust prices. In this case, the contestable market result emerges as a generalization of Bertrand competition to the case where there are economies of scale [Tirole (1988, p. 310)]. However, for most industries, especially those that have typically been associated with the concept of natural monopoly, prices adjust much more quickly than can production capacity and its associated sunk costs. Accordingly, this approach to a contestable market equilibrium does not appear to be of much practical interest either.

A case for price and entry regulation based on a natural monopoly rationale therefore requires both significant increasing returns and long-lived sunk costs that represent a significant fraction of total costs. Indeed, this conclusion reflects a century of economic thinking about monopoly and oligopoly issues, with the development of contestable market theories being an intellectual diversion that, at best, clarifies the important role of sunk costs in theories of monopoly and oligopoly behavior.

Models of "wars of attrition" represent an interesting approach to natural monopoly that allows for increasing returns, sunk costs, exit, textbook monopoly pricing, and no incentives for re-entry in the face of textbook monopoly pricing at the end of the war [e.g. Tirole (1988, p. 311)]. In these models (to simplify considerably) there are two identical firms in the market at time 0. They compete Bertrand (for a random length of time) until one of them drops out of the market because the expected profits from continuing to stay in the market is zero. The remaining firm charges the monopoly price until there is entry by a second firm. However, re-entry by a competing firm is not profitable because the potential entrant sees that post-entry it will have to live through a war of attrition ($p = c$) and, even if it turned out to be the survivor, the expected profits from entry are zero. In this kind of model there is a period of intense competition when

prices are driven to marginal costs.[5] There is also inefficient duplication of facilities during this time period. Then there is a monopoly that "naturally" emerges at some point which charges a textbook pure monopoly price since it is not profitable for an entrant to undercut this price when faced with the threat of a price war. (There remains the question of why both firms entered in the first place.) This kind of war of attrition has been observed repeatedly in the early history of a number of industries that are often considered to have natural monopoly attributes: competing electric power distribution companies, railroad and urban transit lines in the late 19th and early 20th centuries and competing cable TV companies more recently.

War of attrition models also have interesting implications for the kind of "rent seeking" behavior identified by Posner (1975). Monopolies are valuable to their owners because they produce monopoly profits. These potential profits create incentives for firms to expend resources to attain or maintain a monopoly position. These resource expenditures could include things like investments in excess capacity to deter entry, duplication of facilities in the face of increasing returns as multiple firms enter the market to compete to be the monopoly survivor, and expenditures to curry political favor to obtain a legal monopoly through patent or franchise. In the extreme, all of the monopoly rents could be dissipated as a result of these types of expenditures being made as firms compete to secure a monopoly position. The worst of all worlds from a welfare perspective is that all of the monopoly profits are competed away through wasteful expenditures and consumers end up paying the monopoly price.[6]

The combination of increasing returns (and the multiproduct equivalents) combined with a significant component of long-lived sunk costs brings us naturally to more conventional monopoly and oligopoly models involving barriers to entry, entry deterrence and predation [Tirole (1988, Chapters 8 and 9)]. The natural monopoly problem and general models of barriers to entry, entry deterrence and oligopoly behavior are linked together, with natural monopoly being an extreme case. Sunk "capacity" costs create an asymmetry between firms that are "in" the market and potential entrants. This asymmetry can act as a *barrier to entry* by giving the first mover advantage to the firm that is the first to enter the market (the incumbent). Once costs have been sunk by an entrant they no longer are included in the opportunity costs that are relevant to the incumbent firm's pricing decisions. Sunk costs have commitment value because they cannot be reversed. This creates opportunities for an incumbent or first mover to behave strategically to deter entry or reduce the scale of entry.

[5] Since this is a repeated game it is possible that there are dynamic equilibria where the firms tacitly collude and keep prices high or non-cooperative price games with fixed capacity which lead to Cournot outcomes with higher prices.

[6] The war of attrition model that I outlined above is not this bad. There is wasteful "duplication" of facilities prior to the exit of one of the firms but prices are low so consumers benefit during the price war period. After exit consumers must pay the monopoly price, but the costs of duplication are gone. This outcome is worse than the second-best associated with the perfectly contestable market outcome with increasing returns by (effectively) no sunk costs. [Tirole (1988, pp. 311–314)].

In the simplest models of sequential entry with sunk costs and increasing returns [Tirole (1988, pp. 314–323)] firms compete in the long run by making capacity commitments, including how much capacity to accumulate upon entering a market and, for a potential entrant considering to enter to compete with an incumbent, whether or not it will commit capital to support even a modest quantity of capacity needed to enter the market at all. In making this decision the potential entrant must take account of the nature of the competition that will determine prices and entry post entry, at post-entry capacity and output levels. If the incumbent can profitably and credibly make commitments that indicate to the potential entrant that it will be unprofitable to enter due to the nature of the post-entry competition it will face, then competitive entry may be deterred. In these sequential entry games, the presence of sunk costs alone does not generally deter entry, but rather the strategic behavior of the first mover can reduce the amount of capacity the entrant commits to the markets and as a result, sustain post-entry prices above competitive levels and post-entry output below competitive levels. The combination of sunk costs and increasing returns can make small scale entry unprofitable so that the incumbent may deter entry completely.

Joe Bain (1956) characterized alternative equilibria that may arise in the context of significant economies of scale (to which today we would add multiproduct cost complementarities and sunk costs as well) that were subsequently verified in the context of more precise game-theoretic models [Tirole (1988, Chapter 8)]. These cases are:

Blockaded entry: Situations where there is a single firm in the market that can set the pure monopoly price without attracting entry. The incumbent competes as if there is no threat of entry. A situation like this may emerge where economies of scale are very important compared to the size of the market and where sunk costs are a large fraction of total costs. In this case, potential entrants would have to believe that if they entered, the post-entry competitive equilibrium would yield prices and a division of output that would not generate enough revenues to cover the entrant's total costs. This is the classic "pure monopoly" case depicted in microeconomics textbooks.

Entry deterrence: There is still no entry to compete with the incumbent, but the incumbent had to take costly actions to convince potential entrants that entry would be unprofitable. This might involve wasteful investments in excess capacity to signal a commitment to lower post-entry prices or long-term contracts with buyers to limit ("foreclose") the market available for a new entrant profitable to serve (Aghion and Bolton).

Accommodated entry: It is more profitable for the incumbent to engage in strategic behavior that *accommodates* profitable entry but limits the profitability of entry at other than small scale. Here the incumbent sacrifices some short-term pre-entry profits to reduce the scale of entry to keep prices higher than they would be if entry occurred at large scale.

2.6. Empirical evidence on cost subadditivity

Despite the extensive theoretical literature on natural monopoly, there is surprisingly little empirical work that measures the extent to which the costs of producing services that are typically thought of as natural monopolies are in fact subadditive. The most extensive research on the shape of firm level cost functions has been done for electricity [e.g. Christiansen and Greene (1976), Cowing (1974), Joskow and Rose (1985, 1989); Jamasb and Pollitt (2003)]. There has also been empirical work on cost attributes of water companies [Teeples and Glyer (1987)], telecommications firms [Evans (1983), Gasmi, Laffont, and Sharkey (2002)], cable television companies [Crawford (2000)], urban transit enterprises [Gagnepain, P. and M. Ivaldi (2002)], and multi-product utilities [Fraquelli, G., M. Picenza, and D. Vannoni (2004)]. Empirical analysis tends to find economies of scale (broadly defined) out to some level of firm output. However, much of this work fails properly to distinguish between classical economies of scale and what is best thought of as economies of density. Thus, for example, economies of scale in the distribution of natural gas may be exhausted by a firm serving let's say 3 million customers on an exclusive basis in a specific geographic area. However, whatever the size of the geographic area covered by the firm it would still be very costly to run two competing gas distribution systems down the same streets, because there are economies of scale or "density" associated with the installation and size of the pipes running down each street.

3. Why regulate natural monopolies?

It is important to recognize that in reality there is not likely to be a bright line between industries that are "natural monopolies" and those that are (imperfectly) "competitive." Whether an industry is judged to have classical natural monopoly characteristics inevitably depends on judgments about the set of substitute products that are included in the definition of the relevant product market (e.g. are Cheerios and Rice Crispies close enough products to be considered to be in the same product market? Are cable TV and Direct Broadcast Satellite in the same relevant product market?) and the geographic expanse over which the market is regulated (e.g. a supermarket may technically have natural monopoly characteristics if the geographic market is defined very narrowly, but may have no market power since consumers can easily switch between outlets at different geographic locations and the market cannot discriminate between consumers with good substitutes and those without). Moreover, many "competitive" industries are imperfectly competitive rather than perfectly competitive. They may have production technologies that give individual firms economies of scale but there is little cost sacrifice if there are several firms in the market. Or firms may have technologies that exhibit economies of scale over the production of a narrowly defined product or brand but there are many "natural monopolies" producing competing products or brands that are close substitutes for it and constrain the ability of suppliers to exercise market power.

In these cases, competition may be imperfect but the (theoretical) social welfare costs compared to the best *feasible* alternative industry configurations given economies of scale, differentiated product attributes, and break-even constraints may be quite small. This suggests that the technical definitions of natural monopoly employed (normative or positive) must be carefully separated from the questions of whether and how to regulate a particular industry.

The standard normative economic case for imposing price and entry regulations in industries where suppliers have natural monopoly characteristics is that (a) industries with natural monopoly characteristics will exhibit poor economic performance in a number of dimensions and (b) it is feasible in theory and practice for governments to implement price, entry and related supporting regulations in ways that improve performance (net) compared to the economic performance that would otherwise be associated with the unregulated market allocations. That is, the case for government regulation is that there are costly market failures whose social costs (consequences) can in principle be reduced (net) by implementing appropriate government regulatory mechanisms.

This "market failures" case for government regulation naturally leads to four sets of questions. First, what is the nature and magnitude of the performance problems that would emerge absent price and entry regulation in industries with natural monopoly characteristics? Second, what regulatory instruments are practically available to stimulate performance improvements and what are their strengths and weaknesses? Third, what are the performance attributes of the industry configuration that would be expected to emerge in a regulated environment? Fourth, are imperfect regulatory outcomes, on balance, likely to be superior to imperfect market outcomes taking all relevant performance criteria into account, including the direct and indirect costs of government regulation itself?

3.1. Economic efficiency considerations

The *economic efficiency* case for government regulation when an industry has natural monopoly characteristics has focused on a number of presumed attributes and the associated inefficiencies of market outcomes that are thought would arise in the absence of government regulation. Figure 5A displays two potential equilibria for an industry supplied by one single-product firm with subadditive costs. These equilibria provide normative benchmarks against which the performance attributes of "unregulated natural monopoly" can be compared. The firm's costs (AC_e and MC_e) assume that the firm produces a given level of output efficiently given input prices and technology. The price p_o reflects a second-best linear price that allows the firm just to cover its production costs and clears supply and demand. The price p_e is the first-best efficient price ($p = MC$) that leaves the regulated firm with a deficit and therefore requires government subsidies. Note that p_e is efficient in a broader general equilibrium sense only if we ignore the costs the government incurs to raise the revenues required to raise the funds to pay subsidies these through taxation. I will focus here on the case where the

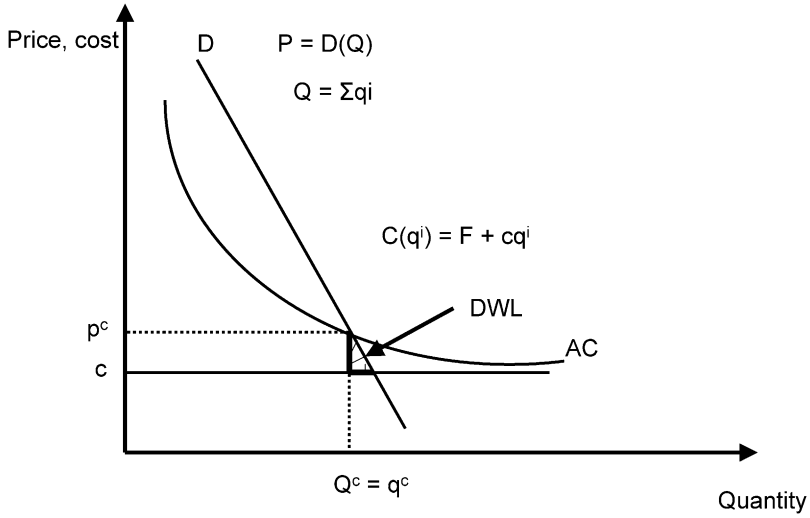


Figure 5A. Break-even price and dead-weight loss.

firm must break-even from the revenues it earns by selling services subject to price regulation to consumers.

Figure 5B depicts an alternative "unregulated natural monopoly" equilibrium where there are sunk costs and barriers to entry. The firm's production costs are now depicted as c_m (to keep the figure from becoming too confused, I have left out the average cost curve AC_M from Figure 5A which we should think of as being higher than AC_e and c_e), reflecting inefficient production by the monopoly, and the price charged by the firm is now $p_m > p_o > p_e$. In Figure 5B the rectangle marked with an "X" depicts the cost or "X-inefficiency" at output level Q_M associated with the monopoly configuration. The firm also spends real resources equal to R per year to maintain its monopoly position, say through lobbying activity or carrying excess capacity to deter entry. The case for regulation starts with a comparison of the attributes of the unregulated natural monopoly equilibrium depicted in Figure 5B with the efficient (first or second best with linear prices) equilibria depicted in Figure 5A.

Inefficient Price Signals: Prices greater than marginal cost: As have seen above, if a single or multiproduct monopoly naturally emerges (and is sustainable) in markets that are "contestable," then the resulting monopoly will not have much market power. At worst, the monopoly will set prices above marginal cost to satisfy a break-even constraint ($p = AC$ in the single product case and under certain conditions Ramsey prices in the multiproduct case (Baumol, Bailey, and Willig, 1977—more on this below). This in turn leads to the standard dead- weight loss triangle associated with the gap between prices and marginal cost (depicted by the triangle marked DWL in Figure 5A). However, these are the second-best linear prices and, assuming that public policy requires

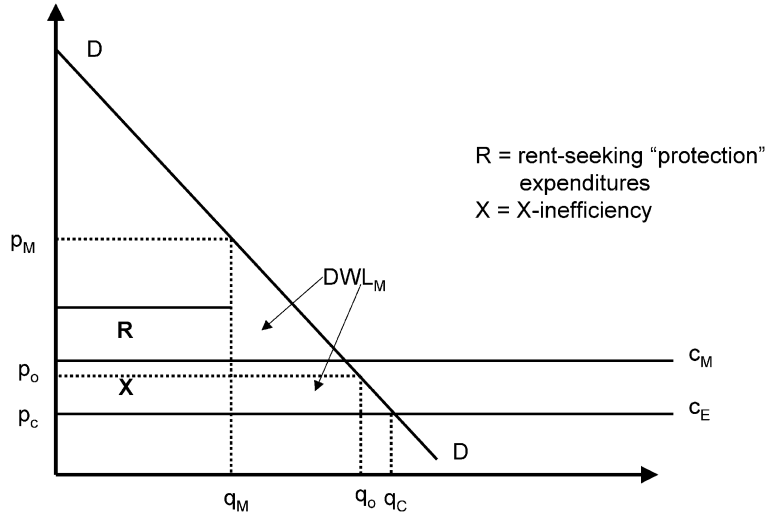


Figure 5B. Potential monopoly inefficiencies.

regulated firms to break-even and to charge linear prices, a regulator could not do any better. This is the second-best price p_o depicted in Figure 5A.

It has been argued that even with contestable markets we could do even better by regulating the monopoly and forcing it to sell at prices equal to marginal costs, using government subsidies to make up the difference between revenues and total costs. This argument normally assumes that the government can raise funds to finance the deficit without incurring any distortionary costs from the tax system put in place to generate the associated government revenues. Since governments do not generally rely on non-distortionary lump sum taxes to raise revenues, the theoretical case for regulating a natural/legal monopoly so as to constrain prices to equal marginal cost must depend on a comparison between the costs of distortions created by prices charged for the regulated services that exceed marginal cost and the costs of distortionary taxes that are otherwise required to pay for the firm's deficit. If the demands for the products and services sold by the regulated firm are fairly inelastic, as is often the case, the distortions resulting from raising prices above marginal cost to balance revenues and costs may not be larger than the distortions caused by increasing taxes to raise the revenues required to close the gap between revenues and costs when prices for the regulated product are force set equal to the relevant marginal costs [Laffont (1999)].

Putting the government subsidy arguments aside for the moment, if one believes that a monopoly has naturally emerged in a setting consistent with the assumptions associated with contestable markets then monopoly price distortions do not create a very good argument for price and/or entry regulation. That is, if the prices in Figure 5B where the same as those in 5A then from a pricing perspective there would be no loss from unregulated natural monopolies.

The more interesting "market failures" case for regulation to mitigate distortions associated with monopoly prices arise in situations in which there are significant barriers to entry and unregulated prices can be sustained at levels far above both marginal cost and average cost. This is the case depicted in Figure 5B where $p_M > p_o$. Since the market power possessed by an incumbent monopoly depends on both the presence of entry barriers and the elasticity of demand for the products sold by the firm, the social costs of monopoly will be higher the more important are entry barriers and the more inelastic is the demand for the relevant products. The polar case is one of blockaded entry (Bain, 1956) where the incumbent dominant firm faces a market demand with elasticity ε_d and sets the monopoly price:

$$P_M = MC/(1 + 1/\varepsilon_d)$$

and the Lerner Index of monopoly power is given by

$$(P_M - MC)/P_M = 1/\varepsilon_d$$

In this case, P_M is the highest price that a monopoly profitably can charge. The incumbent may charge a lower price to accommodate entry or through contracts to deter entry. After entry occurs, prices will likely fall as a result of there being more competition in the market, but they may not fall to the level where total revenues and total costs are equal ($P = AC$). That is, oligopoly price distortions may remain for some period of time. In all of these cases, the firm will charge prices greater than p_o, produce positive ("excess") economic profits that it will have an incentive to invest resources in order to protect, and yield a dead-weight loss from excessive prices alone (area DWL_M in Figure 5B) relative to the dead-weight loss at the break-even uniform unit (linear) price level ($P = AC$ in the single product case). However, if the elasticity of demand is very large in absolute value, any distortion resulting from monopoly pricing will be small.

Inefficient costs of production (including inefficient entry and exit): By definition, a natural monopoly involves production conditions such that it is less costly to produce output in a single firm than in two or more firms. In a contestable markets environment the monopoly in the market has high powered incentives to minimize production costs since it can be replaced instantly by a firm that will to supply at a price equal to average "minimum" (efficient production) total cost. Accordingly, firms or markets that are candidates for regulation must depart from the assumptions associated with contestable markets. That is, we should focus on cases where there are significant scale and scope economies and sunk costs represent a significant fraction of total costs.

In such cases, one potential source of increased production costs arises from the strategic behavior that an incumbent monopoly may engage in order to deter entry and protect its monopoly position. This may entail building excess capacity or spending resources in other ways ("rent seeking" behavior) to obtain or protect a monopoly position. Potentially all of the monopoly profits associated with the pure monopoly outcome may be "wasted" in this way. This type of social cost is depicted as the rectangle marked "R" in Figure 5B.

A second potential source of higher production costs results from inefficient entry of competitors. If the industry has natural monopoly attributes and multiple firms enter the market to supply output—even if competitors eventually exit after a war of attrition—excessive costs are naturally incurred due to duplication of facilities the failure to exploit all available economies of scale. Even in a contestable market the natural monopoly equilibrium may not be sustainable and inefficient entry may occur. The cost of duplicated facilities is not reflected in Figure 5B, but can be conceptualized as being related to the increase in average costs caused by each firm producing at a lower (suboptimal) output level.

A third potential source of production cost inefficiencies is the failure of the incumbent monopoly to minimize production costs—produce efficiently—at the output level it is producing, given technology and input prices. Cost minimization requires that the marginal rate of technical substitution between inputs equal the ratio of their respective prices. If we have a two input production function $q = F(K, L)$ where the rental rates for capital (K) and the wage rate for labor (L) are respectively r and w, then cost minimization at any output level requires that $F_K/F_L = r/w$, where F_K is the marginal product of capital and F_L is the marginal product of labor. Neoclassical profit maximizing monopoly firms minimize costs in this way. However, when there is separation of ownership and management and management gets satisfaction from managerial emoluments and gets disutility from effort, monopoly firms that are insulated from competition may exhibit "X-inefficiency" or managerial slack that leads to higher production costs. There is also some evidence that monopolies are more easily organized by unions which may extract some of the monopoly profits in the form of higher wages ($w_M > w$) [Salinger (1984), Rose (1987), Hendricks (1977)]. If wages are driven above competitive levels this will lead firms inefficiently to substitute capital for labor in production. These costs are depicted as $c_m > c_e$ and the associated social cost is depicted as rectangle marked "X" in Figure 5B.

Product quality and dynamic inefficiencies: Although, the issue has largely been unexplored in the context of natural monopoly *per se*, related literature in industrial organization that examines research and development, adoption of innovations in the production and product dimensions and the choice of product quality suggests that monopoly outcomes are likely to differ from competitive outcomes. Moreover, issues associated with the reliability of service (e.g. outages of the electric power network) and various aspects of the quality of service (e.g. queues for obtaining connections to the telephone network) are significant policy issues in many regulated industries. As a general matter, we know that monopoly will introduce a bias in the selection of quality, the speed of adoption of innovations, and investment in R&D. In simple static models of monopoly the bias turns on the fact that a profit maximizing monopoly looks at the willingness to pay for quality of the marginal consumer while social welfare is maximized by focusing on the surplus achieved by the average consumer (Spence, 1975).

However, the size and magnitude of any quality bias, compared to a social welfare-maximizing norm is ambiguous. The monopoly may supply too much or too little quality or have too little or too much incentive to invest in R&D and adopt innova-

tions depending on the circumstances, in particular whether the incumbent monopoly is threatened by potential entry, as well as the existence and nature of patent protection and spillovers from R&D [Tirole (1988, pp. 100–106, 361–414)]. This is not the place to review the extensive literature on the relationship between market structure and innovation, but I note only that it raises potentially important dynamic efficiency issues with market structures that evolve into monopolies. On the one hand, in situations where there are significant spillovers from R&D and innovation that would otherwise be captured by competing firms and lead to underinvestment in innovation in the product and process dimensions, regulatory policies that facilitate the internalization of these spillover effects, for example, by having a single firm serving the entire sector or providing for the recovery of R&D costs in product prices, might increase social welfare. On the other hand, depending on the circumstances, creating a monopoly and regulating the prices it can charge for new products could increase rather than decrease inefficiencies associated with product quality, R&D, and the adoption of product and process innovations.

Firm viability and breakeven constraints: As I have already noted, if the regulated monopoly is a private firm and there are no government subsidies available to support it, the government may be able to regulate the firm's prices and service quality, but it cannot compel it to supply output to balance supply and demand in the long run if it is unprofitable for it to do so. Accordingly, price and entry regulation also must confront one important set of constraints even in an ideal world where regulators have full information about a firm's cost opportunities, managerial effort levels, and attributes of demand faced by the regulated firm (we discuss regulation with asymmetric information in more detail below). Private firms will only supply goods and services if they expect to at least recover the costs of providing these goods and services. The relevant costs include the costs of materials and supplies, compensation necessary to attract suitable employees and to induce them to exert appropriate levels of effort, the direct cost of capital investments in the enterprise, a return of and on those investments, reflecting the opportunity cost of capital, economic depreciation, taxes, and other costs incurred to provide service. If the process through which regulated prices are set does not lead private firms to expect to earn enough revenues to cover these production and distribution costs the firm will not voluntarily supply the services. Since prices are regulated, supply and demand will not necessarily clear and prices that are set too low will lead to shortages in the short run and/or the long run and the use of non-price rationing to allocate scarce supplies. Accordingly, if we are to rely on regulated private monopolies to provide services, the regulatory process must have a price-setting process that provides the regulated firm with adequate financial incentives to induce them to provide services whose value to consumers exceeds the costs of supplying them.

At this point I will simply refer to this requirement as a breakeven-constraint defined as:

$$\sum_n p_i q_i \geq C(q_1, \ldots, q_{n-1}, q_n)$$

where q_i defines the total output of the different products supplied by the firm (or the same output supplied to different groups of consumers that are charged different prices or a combination of both) and $C(*)$ defines the associated costs. For now, let's think about $C(*)$ as being a static measure of the "efficient" level of costs given any particular output configuration. We will address differences between expected costs and realized costs and issues of cost inefficiency in more detail below.

There is an inherent conflict between the firm viability constraint and efficient pricing when costs are subadditive. Efficient pricing considerations would dictate that prices be set equal to marginal cost. But marginal cost pricing will not produce enough revenues to cover total costs, thus violating the firm viability or break-even constraint.[7] A great deal of the literature on price regulation has focused on responding to this conflict by implementing price structures that achieve the break-even constraint in ways that minimize the efficiency losses associated with departures from marginal cost pricing.

Moreover, because the interesting cases involve technologies where long-lived sunk costs are a significant fraction of total costs, the long-term credibility of regulatory rules plays an important role in convincing potential suppliers that the rules of the regulatory game will in fact fairly compensate them for the sunk costs that they must incur to provide service [Laffont and Tirole (1993, Chapter 10), Armstrong, Cowan, and Vickers (1994, pp. 85–91), Levy and Spiller (1994)]. This is the case because once costs are sunk, suppliers must be concerned that they will be "held-up" by the regulator. That is, once the costs are sunk, the regulator is potentially in a position to lower prices to a point where they cover only avoidable costs, causing the firm that has committed the sunk costs to fail to recover them. As I shall discuss presently, creating a regulatory process and judicial oversight system that constrains the ability of a regulatory agency to hold up a regulated firm in this way has proven to be a central component of regulatory systems that have been successful in attracting adequate investment and associated supplies to the regulated sectors. These "credibility" institutions include legal principles governing the formulas used to set prices and to review "allowable" costs, the structure of regulatory procedures and opportunities for judicial review, as well as de jure and de facto restrictions on competitive entry.

3.2. Other considerations

While this chapter will focus on the economic efficiency rationales for and consequences of the regulation of natural monopolies, we must recognize that the nature and performance of the institutions associated with regulated monopoly in practice reflect additional normative public policy goals and the outcomes of interest group politics.

Income distribution, "essential services," cross-subsidization and taxation by regulation: Although simple conceptualizations of economic efficiency are "indifferent" to

[7] In the single product case declining average cost is a necessary condition for marginal cost pricing to be unprofitable. In the multiproduct case, declining ray average cost is a necessary condition for marginal cost pricing to yield revenues that are less than total costs.

the distribution of surplus between consumers and producers, public policy generally is not. Thus, while the efficiency losses from classical monopoly pricing are measured by a welfare triangle reflecting the loss in the sum of consumers' and producers' surplus from higher prices and lower output, public policy has also been concerned with the transfer of income and wealth associated with the excess profits resulting from monopoly pricing as well. Even ignoring the fact that some of the monopoly profits may be eaten up by wasteful "rent seeking" expenditures, and the difficulties of calculating the ultimate effects on the distribution of income and wealth from monopoly pricing, it is clear that regulatory policy has historically been very concerned with mitigating monopoly profits by keeping prices at a level that roughly reflect the regulated firm's total production costs.

It also is quite clear that several of the industries that have evolved as regulated monopolies produce products access to which has come to be viewed as being "essential" for all of a nation's citizens. I use the term "access" here broadly to reflect both physical access (e.g. "universal service")[8] as well as "affordability" considerations. Electricity, telephone, and clean water services fall in this category. The argument is that absent price and entry regulation, suppliers of these services will not find it economical to expand into certain areas (e.g. rural areas) or if they do will charge prices that are too high given the incomes of the individuals living or firms producing (e.g. farms) in those areas. While there are no clear definitions of what kinds of services are essential, how much is essential, or what are the "reasonable" prices at which such services should be provided, these concepts have clearly played a role in the development of regulatory policies in many countries. This being said, it is hard to argue that food, for example, is any less essential than electricity. Yet there has been no interest in creating regulated legal monopolies for the production and distribution of food. Low-income consumers or residents of rural areas could simply be given subsidies by the government to help them to pay for the costs of services deemed essential by policymakers, as is the case for food stamps. Accordingly, the case for regulated monopoly and the case for subsidies for particular geographic areas or types of consumers appear to be separable policy issues that can in principle be addressed with different policy instruments.

These issues are joined when an industry does have natural monopoly characteristics, and the introduction of government regulation of prices and entry creates opportunities to use the regulated monopoly itself as a vehicle for implementing a product-specific, geographic, customer-type specific internal subsidy program rather than relying on the government's general budget to provide the subsidies directly. With regulated *legal monopoly* that bars competitive entry, regulated prices in some geographic areas or the prices charged to some classes of consumers or for some products can be set at levels above what would prevail if economic efficiency criteria alone were applied to set prices (more on this presently). The excess revenues generated by increasing these

[8] A universal service rationale may also be justified by the desire to internalize network externalities (Katz and Shapiro, 1986). Network externalities may also be a source of cost subadditivity.

prices above their efficient level can then be used to reduce prices to the target classes of customers, leaving the overall level of revenues produced from the menu of regulated prices equal to the total costs incurred by the firm. Richard Posner has referred to this phenomenon as *taxation by regulation* (Posner, 1971) and views government regulation of prices as one instrument of public finance [see also Hausman (1998)].

This phenomenon is also often referred to loosely as *cross-subsidization*. The notion is that one group of consumers subsidizes the provision of service to another group of customers by paying more than it costs to provide them with service while the other group pays less. However, when a firm has natural monopoly characteristics, an objective definition of "cross-subsidization" is not straightforward. When cost functions are subadditive and a natural monopoly is sustainable, break-even prices will generally be above the marginal cost of providing service to any individual or group of consumers. At least some consumers of some products produced by the natural monopoly must pay more than the incremental cost of serving them to satisfy a break-even constraint for the regulated firm. And, as we shall see below, efficient prices will generally vary from customer to customer when marginal cost-based prices do not yield sufficient revenues to cover total costs. Are consumer's paying prices that yield relatively high margins (difference between price and marginal cost) necessarily "subsidizing" consumers paying prices that yield lower margins?

More refined definitions of cross-subsidization have evolved that better reflect the attributes of subadditive cost functions [Sharkey (1982), Faulhaber (1975)]. A price configuration does not involve cross-subsidies (it is "subsidy free") if:

(a) All consumers pay at least the average incremental costs of providing them with service and

(b) No consumers or groups of consumers pay more than the "stand-alone costs" of providing them with service. Stand-alone costs refer to the costs of supplying only one or more groups of consumers that are a subset of the entire population of consumers that demand service at the prices at issue.

If these conditions prevail, consumers who are charged relatively high prices may be no worse off as a consequence of other consumers being charged lower prices and may be made better off than if the latter consumers purchased less (or nothing) from the firm if the prices they are being charged were to increase. This the case because if the contribution to meeting the firm's budget constraint made by the consumers being charged the lower prices is greater than or equal to zero then the remaining consumers will have to pay a smaller fraction of the firm's total costs and are better off than if they had to support the costs of the enterprise on a *stand-alone* basis.

Moreover, if subsidy free prices exist the natural monopoly will also be sustainable [Baumol, Bailey, and Willig (1977), Baumol, Panzar, and Willig (1982)]. On the other hand, if the government endeavors to engage in taxation by regulation in ways that involve setting prices that are not subsidy free, the resulting configuration may not be sustainable. In this case, restrictions on entry—legal monopoly—will be necessary to keep entrants from *cream skimming* the high margin customers away from the incumbent when the stand-alone costs make it profitable to do so [Laffont and Tirole (1990b)].

So, for example, when the U.S. federal government implemented policies in the 1920s to keep regulated local telephone service charges low in order to encourage universal service, subsidize customers in rural areas, etc., it simultaneously kept long-distance prices high to generate enough net revenues from long distance service to cover the costs of the local telephone network that were in excess of local service revenues [Palmer (1992), Crandall and Hausman (2000), Joskow and Noll (1999)]. This created potential opportunities for firms inefficiently to enter the market to supply some of the high-margin long-distance service (the prices were therefore greater than the stand alone costs), potentially undermining the government's ability to utilize taxation by regulation to implement the universal service and income distribution goals. When the costs of creating a competing long distance network were very high, this price structure was sustainable. However, as the costs of long distance telecommunications facilities fell, it became profitable, though not necessarily efficient, for competing entrants to supply, a subset of long distance services: the price structure was no longer sustainable.

Price Discrimination: In the single product case price discrimination involves a firm charging different prices for identical products to different consumers. The discrimination may involve distinguishing between different types of consumers (e.g. residential and commercial customers) and charging different per unit (linear) prices to each group for the same quantities purchased (third-degree price discrimination) or prices may vary depending on the quantities purchased by individual consumers (second-degree price discrimination). In a multiproduct context, price discrimination also encompasses situations where prices are set to yield different "margins" between price and marginal/average incremental cost for different products or groups of products (a form of third-degree price discrimination). The welfare/efficiency consequences of price discrimination by a monopoly in comparison to simple uniform monopoly pricing are ambiguous [Schmalensee (1981)]. Price discrimination could increase or reduce efficiency compared to uniform price-cost margins, depending on the shapes of the underlying demands for the services as well as attributes of the firm's cost function. In a regulated monopoly context, when firms are subject to a breakeven constraint, price discrimination of various kinds can reduce the efficiency losses associated with departures from marginal cost pricing. We will explore these issues presently.

Whatever the efficiency implications of price discrimination, it is important to recognize that real or imagined price discrimination by unregulated monopolies played an important political role in stimulating the introduction of price regulation of "natural monopolies" in the United States. The creation of the Interstate Commerce Commission in 1887 to supervise rail freight rates was heavily influenced by arguments made by shippers served by one railroad that they were being charged much higher prices per mile shipped for similar commodities by the same railroad than where shippers served by competing railroads or with transport alternatives that were close substitutes (e.g. barges) [Kolko (1965), Mullin (2000), Prager (1989a), Gilligan, Marshall, and Weingast (1990)]. Many regulatory statutes passed in the U.S. in the last century have (or had) text saying something like "rates shall be just, reasonable, *and not unduly discriminator*" [Bonbright (1961, p. 22), Clark (1911)]. The development of regulation in the

U.S. has been heavily influenced by the perceived inequities of charging different consumers different prices for what appear to be the same products. When combined with monopoly or very limited competition it has been both a source of political pressure to introduce price regulation and has led to legal and policy constraints on the nature of the price structures that regulatory agencies have at their disposal.

Political economy considerations: By this point it should be obvious that the decision to introduce price and entry regulation, as well as the behavior and performance of regulatory agencies, reflects a broader set of considerations than simply a public interest goal of mitigating the distortions created by unregulated markets with natural monopoly characteristics. Price and entry regulation can and does convey benefits on some groups and impose costs on other groups compared to alternatives, whether these alternatives are no price and entry regulation or alternative mechanisms for implementing price and entry regulation. The potential effects of price and entry regulation on the welfare of different interest groups—different groups of consumers, different groups of suppliers, environmental and other "public interest" groups—has played a significant role in where, when and how price and entry regulation are introduced, when and how regulatory mechanisms are changed, and when and how price and entry regulation may be removed. The nature and magnitude of alternative configurations of price and entry regulation on different interest groups, the costs and benefits these groups face to organize to influence regulatory laws and the behavior of regulatory agencies, and how these groups can use the institutions of government (legislature, executive, judicial) to create regulatory (or deregulatory) laws and influence regulatory behavior and outcomes is a very complex subject. The extensive relevant literature has been reviewed elsewhere (e.g. Noll, 1989) and much of it is covered as well in Chapter 22 (McNollgast) of this handbook. It is not my intention to review it again here. However, there are a number of general lessons learned from this literature that are worth noting as background for the rest of the material in this chapter.

For many years students were taught that regulation had been introduced to respond to natural monopoly problems—a "public interest" view of the introduction of price and entry regulation [Stigler (1971), Posner (1974)]. This view confused the *normative* market failures case for why it might be desirable to introduce price and entry regulation to achieve public interest goals with the *positive* question of why price and entry regulation was actually introduced in a particular industry at a particular time. One cannot and should not assume that because an industry is subject to price and entry regulation it is necessarily a "natural monopoly" in any meaningful sense. The introduction of price and entry regulation and the nature of the regulatory mechanisms used to implement it reflect political considerations that are the outcome of interest group politics [Peltzman (1989)]. There are many industries that have been subject to price and entry regulation (e.g. trucking, oil and natural gas production, various agricultural commodities) where there is no evidence of natural monopoly characteristics or the associated economic performance problems. Because regulation typically involves regulation of both prices and entry, it can be and has been used in some cases to keep prices high rather than low and to restrict competition where it would otherwise lead to lower prices, lower costs, and

other efficiency benefits. Each situation must be judged on the merits based on relevant empirical analysis of firm and industry cost and demand characteristics as well as the effects of regulation on firm behavior and performance.

Whatever the rationale for introducing price and entry regulation, we should not assume that regulatory agencies can and will use the most effective mechanisms for achieving public interest goals that may be available to them. Political considerations driven by interest group politics not only play a role in the introduction of price and entry regulation, but in how it is implemented by regulatory authorities [Weingast and Moran (1983), Noll (1989)]. While policymakers frequently refer to "independent" regulatory agencies in the abstract, the reality is that no regulatory agency is completely independent of political influences. This political influence is articulated by who is appointed to lead regulatory authorities, by legislative oversight and budget control, by the election of commissioners in states with elected commissions, and by the resources that different interest groups can bring to the regulatory process itself [McCubbins (1985), McCubbins, Noll, and Weingast (1987), Joskow, Rose, and Wolfram (1996), Hadlock, Lee, and Parrino (2002)].

Even under the best of circumstances, regulatory institutions can respond effectively to the goals established for them only imperfectly. Regulation leads to direct costs incurred by the agency and those groups who are involved with the regulatory process as well as indirect costs associated with distortions in regulated firm prices, costs, profits, etc., that may result from poorly designed or implemented regulatory mechanisms. The direct costs are relatively small. The indirect costs are potentially very large.

Firms may seek to enter an industry subject to price and entry regulation even if entry is inefficient. This result may flow from political constraints that influence the level and structure of regulated prices and make entry look profitable even though it is inefficient because the regulated price signals are inefficient. Distinguishing between efficient entry requests (e.g. due to technological change, new products, excessive costs of the regulated incumbent) and inefficient entry (e.g. responding to a price structure that reflects significant cross-subsidies) is a significant challenge that requires industry-specific assessments of the presence of natural monopoly characteristics and the distortions that may be caused by inefficient regulation.

3.3. Regulatory goals

Since the focus of this essay is on the economic efficiency rationales for price and entry regulation, the regulatory goals that will guide the design of effective regulatory mechanisms and institutions and against which the performance of regulatory institutions will be evaluated should reflect the same efficiency considerations. In what follows I will focus on the following regulatory goals:

Efficient pricing of goods and services: Regulated prices should provide consumers with efficient price signals to guide their consumption decisions. Ideally, prices will equal the relevant marginal or incremental costs. However, firm-viability and potentially

other constraints will necessarily lead to departures from first-best prices. Accordingly, second-best pricing given these constraints will be the goal on the pricing front.

Efficient production costs: The natural monopoly rationale for restricting entry to a single firm is to make it possible for the firm to exploit all economies of scale and economies of scope that are made feasible by the underlying technology, taking into account the organizational and related transactions costs associated with firms of different horizontal and vertical scales. Textbook presentations of natural monopoly regulation typically take the firm's cost function as given and focus on specification of optimal prices given the firm's costs and break-even constraint. However, by controlling a regulated firm's prices and profits and eliminating the threat of competitive entry, we may simultaneously sharply curtail the incentives that lead competing firms to seek to minimize costs from both static and dynamic perspectives. Moreover, regulation may significantly reduce the efficiency incentives that are potentially created by the market for corporate control by imposing lengthy regulatory review requirements and capturing the bulk of any cost savings resulting from mergers and acquisitions for consumers through lower regulated prices. Regulators need to be focused on creating substitute incentive mechanisms to induce regulated firms to minimize costs by adjusting inputs to reflect the relative input prices, to exert the optimal amounts of managerial effort to control costs, to constrain costly managerial emoluments and other sources of X-inefficiency, and to adopt new process innovations in a timely and efficient manner.

Efficient levels of output and investment (firm participation and firm-viability constraints): The regulated firm should supply the quantities of services demanded by consumers and make the investments in facilities necessary to do so in a timely and efficient manner. If private firms are to be induced to supply efficiently they must perceive that it is privately profitable to do so. Accordingly, regulatory mechanisms need to respect the constraint that private firms will only invest if they expect the investment to be profitable ex ante and will only continue to produce if they can cover their avoidable costs ex post.

Efficient levels of service quality and product variety: Products may be provided with varying levels of service quality and reliability. Different levels of service quality and reliability carry with them different costs. Consumer valuations of service quality and reliability may vary widely as well. Regulators should be concerned that the levels of service quality and reliability, and the variety of quality and reliability options available to consumers reflect consumer valuations and any costs associated with providing consumers with a variety of levels of quality and reliability from which they can choose. Physical attributes of the networks which characterize industries that have often been subject to price and entry regulation may limit the array of product qualities that can be offered economically to consumers. For example, on a typical electric distribution network, individual consumers cannot be offered different levels of network reliability because the physical control of the distribution network is at the "neighborhood" rather than the individual levels (Joskow and Tirole, 2005).

Monopoly profit and rent extraction considerations: While simple models of social welfare (e.g. the sum of consumers' plus producers' surplus) are agnostic about the dis-

tribution of surplus between consumers and producers, it is clear that regulatory policies are not. In addition to the efficiency distortions caused by monopoly pricing, extracting the excess profits associated with monopoly profits for the benefit of consumers is also an important goal of most regulatory laws. It is the flip side of the firm viability constraint. The regulated firm's profits must be "high enough" to induce it to supply efficiently, but "no higher" than is necessary to do so. This goal can be rationalized in a number of ways. I prefer to view it as an articulation of a social welfare function that weights consumers' surplus more than producers' surplus subject to a firm viability or breakeven constraint. Alternatively, one might rationalize it as reflecting a concern that some or all of the monopoly profits will be transformed into wasteful "rent seeking" expenditures by the regulated firm to enable it to retain its monopoly position.

Distributional Goals: To the extent that other income distribution goals (e.g. universal service goals) are assigned to the regulated firm, price and quantity mechanisms should be adopted to achieve these goals at minimum cost.

Ultimately, sound public policy must ask whether the potential improvements in performance along the various performance dimensions discussed above relative to unregulated market outcomes—depicted in a simple fashion in Figures 5A and 5B—are likely to be greater than the direct and indirect costs of government regulatory mechanisms. Accordingly, sensible decisions about whether and how to regulate should consider both the costs of imperfect markets and the costs of imperfect regulation.

4. Historical and legal foundations for price regulation

Government regulation of prices can be traced back at least to the period of the Roman Empire when the emperor established maximum prices for roughly 800 items. These actions found support in the doctrine of "the just price" developed by Church authorities [Phillips (1993 p. 90)]. During the Middle Ages, craft guilds developed which licensed and controlled the individuals who could work in specific occupations. Because these guilds had monopoly control over who could work in particular crafts they were regulated. "The obligation of the guilds was to provide service to anyone who wanted it at reasonable prices. The various crafts were known therefore as 'common carriers,' 'common innkeepers,' 'common tailors' and so forth. Since each craft had a monopoly of its trade, they were closely regulated" [Phillips (1993, p. 90)]. During the 16th century, the French government began to issue Royal charters to trading companies and plantations which gave them special privileges, including monopoly status, and in turn subjected them to government regulation [Phillips (1993, p. 90)]. These charters, analogous to modern franchises, have been rationalized as reflecting efforts by governments to induce private investment in activities that advanced various social goals [Glaeser (1927, p. 201)].

The antecedents of American legal concepts of "public interest" and "public utilities" that were the initial legal foundations for government price and entry regulation can be found in English Common law. "Under the common law, certain occupations or

callings were singled out and subjected to special rights and duties. These occupations became known as 'common callings,'. . . . A person engaged in a common employment had special obligations . . . , particularly the duty to provide, at reasonable prices, adequate services and facilities to all who wanted them" [Phillips (1993, p. 91)]. English common law regulations were carried over to the English colonies and during the Revolution several colonies regulated prices for many commodities and wages [Phillips (1993, pp. 91–92)]. However, after the American Revolution, government regulation of prices and entry faded away as the United States developed a free market philosophy that relied on competition and was hostile to government regulation of prices and entry [Phillips (1993, p. 92)]. Following the Civil War, and especially with the development of the railroads and the great merger wave of the late 1890s, policymakers and the courts began again to look favorably on price and entry regulation under certain circumstances. The Granger Movement of the 1870s focused on pressuring the states and then the federal government to regulate railroad freight rates. State regulation of railroads by special commissions began in the Midwestern states and then spread to the rest of the country [Phillips (1993, p. 93)]. The first federal economic regulatory agency, the Interstate Railroad Commission (ICC), was established in 1888 with limited authority to regulate the structure of interstate railroad rates. This authority was greatly expanded during the first two decades of the 20th century [Gilligan, Marshall, and Weingast (1989), Mullin (2000), Prager (1989a), Kolko (1965), Clark (1911)].

In the U.S., it was widely accepted as a legal matter that a state or municipality (with state authorization) could issue franchises or concessions to firms seeking to provide certain services using rights of way owned by the municipality and to negotiate the terms of the associated contracts with willing suppliers seeking to use such state and municipal rights of way [Hughes (1983), McDonald (1962)]. These firms proposed to use state or municipal property and the state could define what the associated terms and conditions of contracts to use that property would be. However, the notion that a municipal, state or the federal government could on its own initiative independently impose price regulations on otherwise unwilling private entities was a more hotly contested legal issue about which the Supreme Court's views have changed over time.[9]

Until the 1930s, the U.S. Supreme Court was generally fairly hostile to actions by state and federal authorities to restrict the ability of private enterprises to set prices freely without any restrictions imposed by government [Clemens (1950, pp. 12–37)] except under very special circumstances. Such actions were viewed as potentially violating Constitutional protections of private property rights, due process and contracts. On the one hand, the commerce clause (Article I, Section 8, Clause 3) gives the federal government the power to "regulate commerce with foreign nations and among the several states . . . ." On the other hand, the due process clause (Fifth Amendment) and the equal protection of the laws clause (Article Fourteen), and the obligation of contracts clause (Article I, Section 10) restricts the regulatory powers of the government

[9] The relevant Court decisions are discussed in Clemens (1950, pp. 49–54).

[Clemens (1950, pp. 45–48)]. The courts initially recognized some narrow exceptions to the general rule that the government could not regulate prices in light of the protections provided by the Fifth and Fourteenth Amendments; for example when there were emergencies that threatened public health and safety [Bonbright (1961, p. 6)]. And gradually over time the courts carved out additional exceptions "for certain types of business said to have been 'dedicated to a public use' or 'affected with the public interest,'. . ." [Bonbright (1961, p. 6)]. Railroads, municipal rail transit systems, local gas and electricity systems and other "public utilities" became covered by these exceptions.

One would not have to be very creative to come up with a long list of industries that are "affected with the public interest" and where investments had been "dedicated to a public use." And if such vague criteria were applied to define industries that could be subject to price and entry regulation, there would be almost no limit to the government's ability to regulate prices for reasons that go well beyond performance problems associated with natural monopoly characteristics. However, at least up until the 1930s, the courts had in mind a much less expansive notion of what constituted a "public utility" whose prices and other terms and conditions of service could be legitimately regulated by state or federal authorities (or municipal authorities by virtue of power delegated to them by their state government).[10] The two criteria where (a) the product had to be "important" or a "necessity" and (b) the production technology had natural monopoly characteristics [Bonbright (1961, p. 8)]. Clemens (1950, p. 25) argues that "[N]ecessity and monopoly are almost prerequisites of public utility status." One could read this as saying that the combination of relatively inelastic demand for a product that was highly valued by consumers and natural monopoly characteristics on the supply side leading to significant losses in social welfare are a necessary pre-condition for permitting government price and entry regulation. An alternative interpretation is that the "necessity" refers not so much to the product itself, but rather for the "necessity of price and entry regulation" to achieve acceptable price, output and service quality outcomes when industries had natural monopoly characteristics. In either case, until the 1930s, it is clear that the Supreme Court intended that the situations in which government price regulation would be constitutionally permissible were quite narrow.[11]

The conditions under which governments could regulate price, entry and other terms and conditions of service without violating constitutional protections were expanded during the 1930s.[12] Since the 1930s, federal and state governments have imposed

[10] The landmark case is *Munn v. Illinois 94 U.S. 113 (1877)* where the Illinois state legislature passed a law that required grain elevators and warehouses in Chicago to obtain licenses and to charge prices that did not exceed levels specified in the stature. The importance of the grain storage facilities to the grain shipping business in Chicago and that fact that the ownership of the facilities constituted a virtual monopoly were important factors in the Court's decision. See also *Budd v. New York, 143 U.S. 517 (1892)*.

[11] In a series of subsequent cases the Court made it clear that the conditions under which states could regulate prices were narrow. See *German Alliance Insurance Co. v. Lewis 233 U.S. 389 (1914)*, *Wolff Packing Co. v. Court of Industrial Relations, 262 U.S. 522 (1923), Williams v. Standard Oil Co. 278 U.S. 235 (1929)*.

[12] In *Nebbia vs. New York 291 U.S. 502 (1934)* the Supreme Court upheld a New York State law that created a milk control board that could set the maximum and minimum retail prices for milk sold in the State.

price regulation on a wide variety of industries that clearly do not meet the "necessity and natural monopoly" test discussed above—milk, petroleum and natural gas, taxis, apartment rents, insurance, etc.—without violating the Constitution. Nevertheless, the natural monopoly problem, the concept of the public utility developed in the late 19th and early 20th centuries, and the structure, rules and procedures governing state and federal regulatory commissions that are responsible for regulating industries that meet the traditional public utility criteria go hand in hand.

It should also be recognized that just because an industry can as a legal matter be subject to government price and entry regulation does not mean that the owners of the enterprises affected give up their Constitutional protections under the Fifth and Fourteenth amendments. The evolution of legal rules supporting the right of government to regulate prices and entry and impose various obligations on regulated monopolies were accompanied by a parallel set of legal rules that required government regulatory actions to adhere to these constitutional guarantees. This requirement in turn has implications for regulatory procedures and regulatory mechanisms. They must be consistent with the principle that private property cannot be taken by government action without just compensation. This interrelationship between the conditions under which government may regulate prices and the Constitutional protections that the associated rules and procedures must adhere to are very fundamental attributes of U.S. regulatory law and policy. In particular, they have important implications for the incentives regulated firms have to invest in facilities to expand supplies of services efficiently to satisfy the demand for these service whose prices are subject to government regulation [Sidak and Spulber (1997), Kolbe and Tye (1991)].

5. Alternative regulatory institutions

5.1. Overview

There are a variety of organizational arrangements through which prices, entry and other terms and conditions of service might be regulated by one or more government entities. Legislatures may enact statutes that establish licensing conditions, maximum and minimum prices and other terms and conditions of trade in certain goods and services. This was the approach that led to the Supreme Court's decision in *Munn v. Illinois* where prices were regulated by a statute passed by the Illinois legislature. Indeed, the first "public utilities" were created by legislative acts that granted franchises that specified maximum prices and/or profit rates and provide the first examples of rate of return regulation [Phillips (1993, p. 129)]. When changes in supply and demand conditions led to the need for price changes the legislature could, in principle, amend the statute to make these changes. This type of regulation by legislative act was both clumsy and politically inconvenient [McCubbins (1985), Fiorina (1982), McCubbins, Noll, and Weingast (1987), Hughes (1983)].

Governments can also use the terms of the contracts that they issue to firms which require authorization to use public streets and other rights of way to provide service by including in these "franchise contracts" terms and conditions specifying prices and how they can be adjusted over time [McDonald (1962), Hughes (1983)]. The sectors that are most often categorized as "public utilities" typically began life as local companies that received franchises from the individual municipal governments to whose streets and rights of way they required access to provide service. City councils and agencies negotiated and monitored the associated franchise contracts and were effectively the regulators of these franchisees. However, as contracts, the ability of the municipality to alter the terms and conditions of the franchise agreement without the consent of the franchisee was quite limited [National Civic Federation (1907)]. Most gas, electric, telephone, water and cable TV companies that provide local service and use municipal streets and rights of way still must have municipal franchises, but these franchises typically are little more than mechanisms to collect fees for the use of municipal property as state and federal laws have transferred most regulation of prices and entry to state and/or federal regulatory agencies. The strengths and weaknesses of municipal franchise contracts allocated through competitive bidding are discussed further below.

The "independent" regulatory commission eventually became the favored method for economic regulation in the U.S. at both the state and federal levels [Clemens (1950, Chapter 3), Kahn (1970, p. 10), Phillips (1993, Chapter 4)]. Independent regulatory commissions have been given the responsibility to set prices and other terms and conditions of service and to establish rules regarding the organization of public utilities and their finances. This approach creates a separate board or commission, typically with a staff of engineers, accountants, finance specialists and economists, and gives it the responsibility to regulate prices and other terms and conditions of services provided by the companies that have been given charters, franchises, licenses or other permissions to provide a specific service "in the public interest." The responsibilities typically extend to the corporate forms of the regulated firms, their finances, the lines of business they may enter and their relationships with affiliates. Regulatory agencies are also given various authorities to establish accounting standards and access to the books, records and other information relevant for fulfilling their regulatory responsibilities, to approve investment plans and financings, and to establish service quality standards. Regulated firms are required to file their schedules of prices or "tariffs" with the regulatory commission and all eligible consumers must be served at these prices. Changes in price schedules or tariffs must be approved by the regulatory agency. We will discuss commission regulation in more detail presently.

A final approach to "the natural monopoly problem" has been to rely on public ownership. Under a public ownership model, the government owns the entity providing the services, is responsible for its governance, including the choice of senior management, and sets prices and other terms and conditions. Public ownership may be affected through the creation of a bureau or department of the municipal or state government that provides the services by creating a separate corporate entity organized as a public benefit corporation with the government as its sole owner. In the latter case, the state-

owned company will typically then be "regulated" by a municipal or state department which will approve prices, budgets and external financing decisions. In the U.S. there has been only limited use of public ownership as a response to the natural monopoly problem. The primary exceptions are electricity where roughly 20% of the electricity distributed or generated in the U.S. is accounted for by municipal or state public utility districts (e.g. Los Angeles Department of Water and Power) or federal power marketing agencies (e.g. TVA) and the public distribution of water where state-owned enterprises play a much larger role. Natural gas transmission and distribution, telephone and related communications, and cable television networks are almost entirely private in the U.S. This has not been the case in many other countries in Europe, Latin America, and Asia where state-owned enterprises dominated these sectors until the last decade or so.

There is a long literature on public enterprise and privatization that covers both traditional natural monopoly industries and other sectors where public enterprise spread [e.g. Vickers and Yarrow (1991), Armstrong, Cowan, and Vickers (1994), Megginson and Netter (2001)]. The literature covers price regulation as well as many other topics related to the performance of state-owned utilities. I will not cover the literature on public enterprise or privatization in this essay.

5.2. Franchise contracts and competition for the market

When the supply of a good or service has natural monopoly characteristics "*competition within the market*" will lead to a variety of performance failures as discussed above. While "competition within the market" may lead to these types of inefficiencies, Harold Demsetz (1968) suggested that "*competition for the market*" could rely on competitive market processes, rather than regulation, to select the most efficient supplier and (perhaps) a second-best break-even price structure. The essence of the Demsetz proposal is to use competitive bidding to award monopoly franchise contracts between a government entity and the supplier, effectively to try to replicate the outcomes that would emerge in a perfectly contestable market. The franchise could go to the bidder that offers to supply the service at the lowest price (for a single product monopoly) or the most efficient (second-best) price structure. The franchising authority can add additional normative criteria to the bidding process. Whatever the criteria, the idea is that the power of competitive markets can still be harnessed at the ex ante franchise contract execution stage even though ex post there is only a single firm in the market. Ex post, regulation effectively takes place via the terms and conditions of the contract which are, in turn, determined by competitive bidding ex ante.

For a franchise bidding system to work well there must, at the very least, be an adequate number of ex ante competitors and they must act independently (no collusion). In this regard, one cannot presume that ex ante competition will be perfect competition due to differences among firms in access to productive resources, information and other attributes. Competition among two or more potential suppliers may still be imperfect. The efficiency and rent distribution attributes of the auction will also depend on the specific auction rules used to select the winner and the distribution of information about

costs and demand among the bidders [Klemperer (2002)]. And, of course, the selection criteria used to choose the winner may be influenced by the same kinds of political economy considerations noted above.

More recent theoretical developments in auction theory and incentive theory lead to a natural bridge between franchise bidding mechanisms and incentive regulation mechanisms, a subject that we will explore in more detail below. Laffont and Tirole (1993, Chapter 7) show that the primary benefit of the optimal auction compared to the outcome of optimal regulation with asymmetric information in this context is that competition lowers the prices (rents) at which the product is supplied. In addition, as is the case for optimal regulation with asymmetric information (more below) the franchise contract resulting from an optimal auction is not necessarily a fixed price contract but rather a contract that is partially contingent on realized (audited) costs. The latter result depends on the number of competitors. As the number of competitors grows, the result of the optimal auction converges to a fixed price contract granted to the lowest cost supplier, who exerts optimal effort and leaves no excess profits on the table [Laffont and Tirole (1993, p. 318)]. Armstrong and Sappington (2003a, 2003b) show (proposition 14) that the optimal franchise auction in a static setting with independent costs has the following features: (a) The franchise is awarded to the firm with the lowest costs; (b) A high-cost firm makes zero rent; (b) the rent enjoyed by a low-cost firm that wins the contest decreases with the number of bidders; (c) the total expected rent of the industry decreases with the number of bidders; (d) the prices that the winning firm charges do not depend on the number of bidders and are the optimal prices in the single-firm setting. That is, in theory, with a properly designed auction and a large number of competitors, the outcome converges to the one suggested by Demsetz.

The Demsetz proposal and the related theoretical research seems to be most relevant to natural monopoly services like community trash collection or ambulance services where assets are highly mobile from one community to another (i.e. minimal location-specific sunk costs), the attributes of the service can be easily defined and suppliers are willing to offer services based on a series of repeated short-term contracts mediated through repeated use of competitive bidding. That is, it is most relevant to market environments that are closer to being contestable. It ignores the implications of significant long-lived sunk costs, asymmetric information between the incumbent and non-incumbent bidders, strategic actions changing input prices, changing technology, product quality and variety issues, and incomplete contracts.

As Williamson (1976) has observed, these attributes of the classical real-world natural monopoly industries make once-and-for-all long-term contracts inefficient and not credible. One alternative is to rely on repeated fixed-price short-term contracts. But in the presence of sunk costs and asymmetric information, repeated fixed-price auctions for short-term franchise contracts lead to what are now well known ex ante investment and ex post adaptation problems associated with incomplete complex long-term contracts and opportunistic behavior by one or both parties to the franchise agreement [Williamson (1985)]. Where sunk costs are an important component of total costs, repeated auctions for short-term fixed-price contracts are unlikely to support efficient

investments in long-lived assets and efficient prices for the associated services. This in turn leads to the need for an institutional mechanism to adjudicate contractual disputes. This could be a court or a government agency created by the government to monitor contractual performance, to negotiate adjustments to the franchise contract over time, and to resolve disputes with the franchisee. Goldberg (1976) argues that in these circumstances the franchising agency effectively becomes a regulatory agency that deals with a single incumbent to enforce and adjust the terms of its contract. Joskow and Schmalensee (1986) suggest that government regulation is productively viewed from this contract enforcement and adjustment perspective. For the kinds of industries that are typically thought of a regulated natural monopolies, the complications identified by Williamson and Goldberg are likely to be important.

5.3. *Franchise contracts in practice*

In fact, franchise bidding for natural monopoly services is not a new idea but a rather old idea with which there is extensive historical experience. Many sectors with (arguably) natural monopoly characteristics in the U.S., Europe, Canada and other countries that started their lives during the late 19th and early part of the 20th century, started off life as suppliers under (typically) municipal franchise contracts that were issued through some type of competitive bidding process [Phillips (1993, pp. 130–131), Hughes (1983, Chapter 9)]. The franchise contracts were often exclusive to a geographic area, but in many cases there were multiple legal (and illegal) franchisees that competed with one another [Jarrell (1978), McDonald (1962)] in the same geographic area.

In many cases the initial long-term contracts between municipalities and suppliers broke down over time as economic conditions changed dramatically and the contracts did not contain enforceable conditions to adapt prices, services, and quality to changing conditions, including competitive conditions, and expectations changed [Hughes (1983), McDonald (1962)]. The historical evolution is consistent with the considerations raised by Williamson and Goldberg.[13] Municipal corruption also played a role, as did wars of attrition when there were competing franchises and adverse public reaction to multiple companies stringing telephone and electric wires on poles and across city streets and disruptions caused by multiple suppliers opening up streets to bury pipes and wires [McDonald (1962), National Civic Federation (1907)]. Utilities with municipal franchises began to expand to include many municipalities, unincorporated areas of the state and to cross state lines [Hughes (1983)]. These expansions reflect further exploitation of economies of scale, growing demand for the services as costs and prices fell due to economies of scale, economies of density, technological change, and extensive merger and acquisition activity. Municipalities faced increasing difficulties in regulat-

[13] Though municipal franchise contracts for cable TV service appear not to have had the significant performance problems identified by Williamson (1976). See [Prager (1989b), Zupan (1989a, 1989b)] while federal efforts to regulate cable TV prices have encountered significant challenges [Crawford (2000)].

ing large corporate entities that provided service in many municipalities from common facilities [National Civic Federation (1907), Hughes (1983)]. By around the turn of the 20th century, problems associated with the governance of municipal franchise contracts and their regulation led progressive economists like John R. Commons to favor replacing municipal franchise contracting and municipal regulation with state regulation by independent expert regulatory agencies that could be better insulated from interest group politics generally [McDonald (1962)] and have access to better information and relevant expertise to more effectively determine reasonable prices, costs, service quality benchmarks, etc. [Prager (1990)].

5.4. Independent "expert" regulatory commission

5.4.1. Historical evolution

Prior to the Civil War, several states established special commissions or boards to collect information and provide advice to state legislatures regarding railroads in their states. These commissions were advisory and did not have authority to set prices or other terms and conditions of service [Phillips (1993, p. 132), Clemens (1950, p. 38)]. The earliest state commissions with power over railroad rates were established by "the Granger laws" in several Midwestern states in the 1870s.[14] These commissions had various powers to set maximum rates, limit price discrimination and to review mergers of competing railroads. By 1887, twenty-five states had created commissions with various powers over railroad rates and mergers and to assist state legislatures in the oversight of the railroads [Phillips (1993, p. 132)]. In 1887, the federal government created the Interstate Commerce Commission (ICC) to oversee and potentially regulate certain aspects of interstate railroad freight rates, though the ICC initially had limited authority and shared responsibilities with the states. [Clemens (1950, p. 40)]. The ICC's regulatory authority over railroads was expanded considerable during the first two decades of the 20th century [Mullin (2000), Prager (1989a), Kolko (1965), Gilligan, Marshall, and Weingast (1989)] and was extended to telephone and telegraph (until these responsibilities were taken over by the Federal Communications Commission in 1934) and to interstate trucking in 1935 and domestic water carriers in 1940.

State commission regulation of other "public utility" sectors spread much more slowly as they continued to be subject to local regulation through the franchise contract and renewal process. Massachusetts established the Board of Gas Commissioners in 1885 which had power to set maximum prices and to order improvements in service [Clemens (1950, p. 41)]. Its power was extended to electric light companies two years later. However, the transfer of regulatory power from local governments to state commissions began in earnest in 1907 when New York and Wisconsin created state commissions with jurisdiction over gas distribution, electric power, water, telephone and

[14] Earlier state railroad commissions had fact finding and advisory roles.

telegraph service prices. By 1920 more than two-thirds of the states had created state public utility commissions [Stigler and Friedland (1962), Phillips (1993, p. 133), Jarrell (1978)], a very rapid rate of diffusion of a new form of government regulatory authority, and today all states have such commissions. The authority of the early commissions over the firms they regulated was much less extensive than it is today, and their legal authorities, organization and staffing evolved considerably over time [Clemens (1950, p. 42)].

Federal commission regulation expanded greatly during the 1930s with the Communications Act of 1934 and the associated creation of the Federal Communications Commission (FCC) with authority over the radio spectrum and interstate telephone and telegraph rates, the expansion of the powers of the Federal Power Commission (FPC, now the Federal Energy Regulatory Commission or FERC) by the Federal Power Act of 1935 to include interstate sales of wholesale electric power and transmission service, and interstate transportation and sales of natural gas to gas distribution companies and large industrial consumers, the passage of the Public Utility Holding Company Act of 1935 which gave the new Securities and Exchange Commission (SEC) regulatory responsibilities for interstate gas and electric public utility holding companies, the expansion of the ICC's authority to regulate rates for interstate freight transportation by trucks in 1935, and the creation of the Civil Aeronautics Board (CAB) to regulated interstate air fares in 1938.

It is hard to argue that the growth of federal regulation at this time reflected a renewed concern about performance problems associated with "natural monopolies." The expansion of federal authority reflected a number of factors: the general expansion of federal authority over the economy during the Great Depression and in particular the popularization of views that "destructive competition" and other types of market failure were a major source of the country's economic problems; efforts by a number of industries to use federal regulatory authority to insulate themselves from competition, especially in the transportation areas (railroads, trucks, airlines); as well as the growth of *interstate* gas pipelines, electric power networks, and telephone networks that could not be regulated effectively by individual states.

5.4.2. Evolution of regulatory practice

It became clear to students of regulation and policymakers that effective regulation by the government required expertise in areas such as engineering, accounting, finance, and economics. Government regulators also needed information about the regulated firms' costs, demand, investment, management, financing, productivity, reliability and safety attributes to regulate effectively. Powerful interest groups were affected by decisions about prices, service quality, service extensions, investment, etc. and had incentives to exert any available political and other influence on regulators. The regulated firms and larger industrial and commercial consumer groups were likely to be well organized to exert this kind of influence, but residential and small commercial consumers were likely to find it costly and difficult to organize to represent their interests effectively

through the same political processes. At the same time, the industries subject to regulation were capital intensive, incurred significant sunk costs associated with investments in long-lived and immobile assets and were potentially subject to regulatory hold-ups. The threat of such hold-ups would reduce or destroy incentives to make adequate investments to balance supply and demand efficiently.

The chosen organizational solution to this web of challenges for price and entry regulation in the U.S. during most of the 20th century was the independent regulatory commission [Phillips (1993)]. The commission would have a quasi-judicial structure that applied transparent administrative procedures to establish prices, review investment and financing plans, and to specify and monitor other terms and conditions of service. At the top of the commission would be three to seven public utility "commissioners" who were responsible for voting "yes" or "no" on all major regulatory actions. In most jurisdictions the commissioners are appointed by the executive (governor or the President) and approved by the legislature. They are often appointed for fixed terms and sometimes for terms that are coterminous with the term of the governor. At the federal level and in a number of states no more than a simple majority of the commissioners can be registered in the same political party. In about a dozen states the public utility commissioners are elected by popular vote [Joskow, Rose, and Wolfram (1996)].

Underneath the commissioners is a commission staff which consists of professionals with training in engineering, accounting, finance, and economics and often a set of administrative law judges who are responsible for conducting public hearings and making recommendations to the commissioners. The composition and size of commission staffs varies widely across the states. Commissions adopt uniform systems of accounts and require regulated firms to report extensive financial and operating data to the commission on a continuing basis consistent with these accounting and reporting protocols. Each commission adopts a set of administrative procedures that specifies how the commission will go about making decisions. These procedures are designed to give all interest groups the opportunity to participate in hearings and other administrative procedures, to make information and decisions transparent, and generally to provide due process to all affected interest groups. These procedures include rules governing private meetings between groups that may be affected by regulatory commission proceedings (so-called ex parte rules), rules about the number of commissioners who may meet together privately, and various "sunshine" and "open meeting" rules that require commissioners to make their deliberations public. Regulatory decisions must be based on a reasonable assessment of the relevant facts in light of the agency's statutory responsibilities. Prices must be "just, reasonable and not unduly discriminatory," insuring the consumers are charged no more than necessary to give the regulated firms a reasonable opportunity to recover efficiently incurred costs, including a fair rate of return of and on their investments [*Federal Power Commission v. Hope Natural Gas* 320 U.S. 591 (1944)].

In light of the evolution of constitutional principles governing economic regulation, providing adequate protection for the investments made by regulated firms in assets dedicated to public use plays an important role in the regulatory process and has important implications for attracting investments to regulated sectors. Not surprisingly, these

administrative procedures have evolved considerably over time, with the general trend being to provide more opportunities for interest group participation, more transparency, and fewer opportunities for closed-door influence peddling [Chapter 22, McNollgast (in this handbook)]. Regulatory decisions may be appealed to state or federal appeals courts.

Of course this idealized vision of the independent regulatory commission making reasoned decisions based on an expert assessment of all of the relevant information available often does not match the reality very well. No regulatory agency can be completely independent of political influences. Commissioners and senior staff members are political appointments and while they cannot be fired without just cause they are also unlikely to be appointed or reappointed if their general policy views are not acceptable to the executive or the public (where commissioners are elected). Regulatory agencies are also subject to legislative oversight and their behavior may be constrained through the legislative budgetary process [Weingast and Moran (1983)]. Regulators may have career ambitions that may lead them to curry favor with one interest group or another [Laffont and Tirole (1993, Chapter 16)]. Staffs may be underfunded and weak. Reporting requirements may not be adequate and/or the staff may have inadequate resources properly to analyze data and evaluate reports submitted by the parties to regulatory proceedings. Ex parte rules may be difficult to enforce. The administrative process may be too slow and cumbersome to allow actions to be taken in a timely way. Under extreme economic conditions, regulatory principles that evolved to protect investments in regulated enterprises from regulatory expropriation come under great stress [Joskow (1974), Kolbe and Tye (1991), Sidak and Spulber (1997)]. On the other hand, both the executive branch and the legislature may find it politically attractive to devolve complicated and controversial decisions to agencies that are both expert and arguably independent [McCubbins, Noll, and Weingast (1987)].

All things considered, the performance of the U.S. institution of the independent expert regulatory agency turns on several attributes: a reasonable level of independence of the commission and its staff from the legislative and executive branches supported by detailed due process and transparency requirements included in enforceable administrative procedures, the power to specify uniform accounting rules and to require regulated firms to make their books and operating records available to the commission, a professional staff with the expertise and resources necessary to analyze and evaluate this information, constitutional protections against unreasonable "takings" of investments made by regulated firms, and the opportunity to appeal regulatory decisions to an independent judiciary.

6. Price regulation by a fully informed regulator

Much of the traditional theoretical literature on price regulation of natural monopolies assumes that there is a legal monopoly providing one or more services and a regulatory agency whose job it is to set prices. The regulated firm has natural monopoly characteristics (generally economies of scale in its single product and multiproduct variations)

and the firm is assumed to minimize costs given technology, input prices and output levels (i.e., no X-inefficiency). That is, the firm's cost function is taken as given and issues of production inefficiency are ignored. In the presence of scale economies, marginal cost pricing will typically not yield sufficient revenues to cover total cost. Fully efficient pricing is typically not feasible for a private firm that must meet a break-even constraint in the presence of economies of scale (even with government transfers since government taxation required to raise revenues to transfer to the regulated firm creates its own inefficiencies). Accordingly, the traditional literature on price regulation of natural/legal monopolies focused on normative issues related to the development of second-best pricing rules for the regulated firm given a break-even constraint (or given a cost of government subsidies that ultimately rely on a tax system that also creates inefficiencies). A secondary focus of the literature has been on pricing of services like electricity which are non-storable, have widely varying temporal demand, have high capital intensities and capital must be invested to provide enough capacity to meet the peak demand—the so-called peak-load or variable-load pricing (PLP) problem.

The traditional literature on second-best pricing for natural monopolies assumes that the regulator is fully informed about the regulated firm's costs and knows as much about the attributes of the demand for the services that the firm supplies as does the regulated firm. The regulator's goal is to identify and implement normative pricing rules that maximize total surplus given a budget constraint faced by the regulated firm. Neither the regulated firm nor the regulator acts strategically. This literature represents a normative theory of what regulators *should* do if they are fully informed. It is not a positive theory of what regulators or regulated firms actually do in practice. (Although there is a sense of "normative as positive theory of regulation" in much of the pre-1970s literature on price regulation.)

6.1. Optimal linear prices: Ramsey-Boiteux pricing

In order for the firm with increasing returns to break-even it appears that the prices the firm charges for the services it provides will have to exceed marginal cost. One way to proceed in the single product context is simply to set a single price for each unit of the product equal to its average cost (p_{AC}). Then the expenditures made by each consumer i will be equal to $E_i = p_{\mathrm{AC}} q_i$. In this case p_{AC} is a uniform linear price schedule since the firm charges the same price for each unit consumed and each consumer's expenditures on the product varies proportionately with the output she consumes. In the multiproduct context, we could charge a uniform price per unit for each product supplied by the firm that departs from its marginal or average incremental cost by a common percentage mark-up consistent with meeting the regulated firm's budget constraint. Again the prices charged for each product are linear in the sense that the unit price for each product is a constant and yields a linear expenditure schedule for consumers of each product.

The first question to address is whether, within the class of linear prices, we can do better than charging a uniform price per unit supplied that embodies an equal mark-up over marginal cost to all consumers for all products sold by the regulated firm?

Alternatively, can we do better by engaging in third degree price discrimination, in the case of a single product firm, by charging different unit prices to different types of consumers (e.g. residential and industrial and assuming that resale is restricted) or in the case of multiproduct firms by charging a constant unit price for each product but where each unit price embodies a different markup over its incremental cost?

Following Laffont and Tirole (2000, p. 64), the regulated firm produces n products whose quantities supplied are represented by the vector $\mathbf{q} = (q_1, \ldots, q_n)$. Assume that the demand functions for the price vector $\mathbf{p} = (p_1, \ldots, p_n)$ are $q_k = D_k(p_1, \ldots, p_n)$. The firm's total revenue function is then $R(\mathbf{q}) = \sum_{(i=1,n)} p_k q_k$. Let the firm's total cost function be $C(\mathbf{q}) = C(q_1, \ldots, q_n)$ and denote the marginal cost for each product k as $C_k(q_1, \ldots, q_n)$.

Let $S(q)$ denote the gross surplus for output vector $\mathbf{q}$ with $\frac{\partial S}{\partial q_k} = p_k$. The Ramsey-Boiteux pricing problem [Ramsey, 1927, Boiteux, 1971 (1956)] is then to find the vector of constant unit (linear) prices for the n products that maximizes net social surplus subject to the regulated firm's break-even or balanced budget constraint:

$$\max_q \big\{S(\mathbf{q}) - C(\mathbf{q})\big\} \quad \text{subject to} \tag{1}$$

$$R(q) - C(q) \geq 0 \tag{2}$$

or equivalently, maximizing the firm's profit subject to achieving the Ramsey-Boiteux level of net social surplus:

$$\max_q \big\{R(\mathbf{q}) - C(\mathbf{q})\big\} \quad \text{subject to} \tag{3}$$

$$S(q) - C(q) \geq S(\mathbf{q}^*) - C(\mathbf{q}^*) \tag{4}$$

Where $\mathbf{q}^*$ represent the Ramsey-Boiteux levels of output.

Let $1/\lambda$ represent the shadow price of the constraint in the second formulation above. Then the first order condition for the maximization of (3) subject to (4) for each q_k is given by:

$$\lambda\left(p_k - c_k + \sum_{j=1}^{n} \frac{\partial p_j}{\partial q_k} q_j\right) + p_k - c_k = 0 \tag{5}$$

When the demands for the products produced by the regulated firm are *independent* this reduces to:

$$\frac{p_k - c_k}{p_k} = \frac{\lambda}{1+\lambda}\frac{1}{\eta_k} \tag{6}$$

for all products $k = 1, \ldots, n$ and where η_k is the own-price elasticity of demand for product k. This is often referred to as the *inverse elasticity rule* [Baumol and Bradford (1970)]. Prices are set so that the difference between a product's price and its marginal cost varies inversely with the elasticity of demand for the product. The margin is higher for products that have less elastic demands than for products that have more elastic demand (at the equilibrium prices).

When the products produced by the regulated firm are not independent—they are substitutes or complements—the own-price elasticities in (6) must be replaced with "super-elasticities" that reflect the cross-price effects as well as own-price effects. If the products are substitutes, the Ramsey-Boiteux prices are higher than would be implied by ignoring the substitution effect (the relevant superelasticity is less elastic than the own-price elasticity of good k) and vice versa.

Note that Ramsey-Boiteux prices involve third-degree price discrimination that results in a set of prices that lie between marginal cost pricing and the prices that would be set by a pure monopoly engaging in third-degree price discrimination. For example, rather than being different products, assume that q_1 and q_2 are the same product consumed by two groups of consumers who have different demand elasticities (e.g. residential and industrial consumers) and that resale can be blocked, eliminating the opportunity to arbitrage away differences in prices charged to the two groups of consumers. Then the price will be higher for the group with the less elastic demand despite the fact that the product and the associated marginal cost of producing it are the same. Note as well that the structure, though not the level, of the Ramsey-Boiteux prices is the same as the prices that would be charged by an unregulated monopoly with the opportunity to engage in third-degree price discrimination.

6.2. *Non-linear prices: simple two-part tariffs*

Ramsey-Boiteux prices are still only second-best prices because the per unit usage prices are not equal to marginal cost. The distortion is smaller than for uniform ($p = \text{AC}$ in the single product case) pricing since we are taking advantage of differences in the elasticities of demand for different types of consumers or different products to satisfy the budget constraint yielding a smaller dead-weight loss from departures from marginal cost pricing. That is, there is still a wedge between the price for a product and its marginal cost leading to an associated dead-weight loss. The question is whether we can do better by further relaxing the restriction on the kinds of prices that the regulated firm can charge? Specifically, can we do better if we were to allow the regulated firm to charge a "two-part" price that includes a non-distortionary uniform fixed "access charge" (F) and then a separate per unit usage price (p). A price schedule or tariff of this form would yield a consumer expenditure or outlay schedule of the form:

$$T_i = F + pq_i$$

Such a price schedule is "non-linear" because the average expenditure per unit consumed T_i/q_i is no longer constant, but falls as q_i increases. We can indeed do (much) better from an efficiency perspective with two-part prices than we can with second-best (Ramsey-Boiteux) linear prices (Brown and Sibley, 1986, pp. 167–183).

Assume that there are N identical consumers in the market each with demand $q_i = d(p)$ and gross surplus of S_i evaluated at $p = 0$. The regulated firm's total cost function is given by $C = f_o + cq$. That is, there is a fixed cost f_o and a marginal cost c. Consider a tariff structure that requires each consumer to pay an access charge $A = f_o/N$ and

then a unit charge $p = c$. Consumer i's expenditure schedule is then:

$$T_i = A + pq_i = f_o/N + cq_i$$

This two-part tariff structure is first-best (ignoring income effects). On the margin, each consumer pays a usage price equal to marginal cost and the difference between the revenues generated from the usage charges and the firm's total costs are covered with a fixed fee that acts as a lump sum tax. As long as $A < (S_i - pq_i)$ then consumers will pay the access fee and consume at the efficient level. If $A > (S_i - pq_i)$ then it is not economical to supply the service at all because the gross surplus is less than the total cost of supplying the service (recall S_i is the same for all consumers and $p_i = c$).

Two-part tariffs provide a neat solution to the problem of setting efficient prices in this context when consumers are identical (or almost identical) or A is very small compared to the net surplus retained by consumers (i.e. after paying $pq_i = cq_i$). However, in reality consumers may have very different demands for the regulated service and A may be large relative to $(S_i - cq_i)$ for at least some consumers. In this case, if a single access fee $A = f_o/N$ is charged consumers with relatively low valuations will choose not to pay the access fee and consumer zero units of the regulated service even though their net surplus exceeds cq_i and they would be willing to make at least some contribution to the firm's fixed costs. A uniform two-part tariff would not be efficient in this case. However, if the regulator were truly fully informed about each consumer's individual demand and could prevent consumers from reselling the service, then a "discriminatory" two-part tariff could be tailored to match each consumer's valuation. In this case the customized/access fee A_i charged to each consumer would simply have to satisfy the condition $A_i < (S_i - cq_i)$ and there will exist at least one vector of A_i values that will allow the firm to satisfy the break-even constraint as long as it is efficient to supply the service at all.

If any of the conditions are met for two-part tariffs to be an efficient solution, the welfare gains compared to Ramsey-Boiteux pricing are likely to be relatively large [Brown and Sibley (1986, Chapter 7)].

6.3. *Optimal non-linear prices*

In reality, consumers are likely to be quite diverse and the regulator will not know each individual consumer's demand for the services whose prices they regulate. Can we use a variant of two-part tariffs to realize efficiency gains compared to either Ramsey-Boiteux prices or uniform two-part tariffs? In general, we can do better with non-linear pricing than with simple Ramsey-Boiteux pricing as long as the regulator is informed about the distribution of consumer demands/valuations for the regulated service in the population.

Consider the case where there are two types of consumers, one type (of which there are n_1 consumers) with a "low demand" and another type (of which there are n_2 consumers) with a "high demand." The inverse demand functions for representative type 1 and type 2 consumers are depicted in Figure 6 as $p = d_1(q_1)$ and $p = d_2(q_2)$. The cost function is as before with marginal cost $= c$. If we charge a uniform unit price of $p = c$,

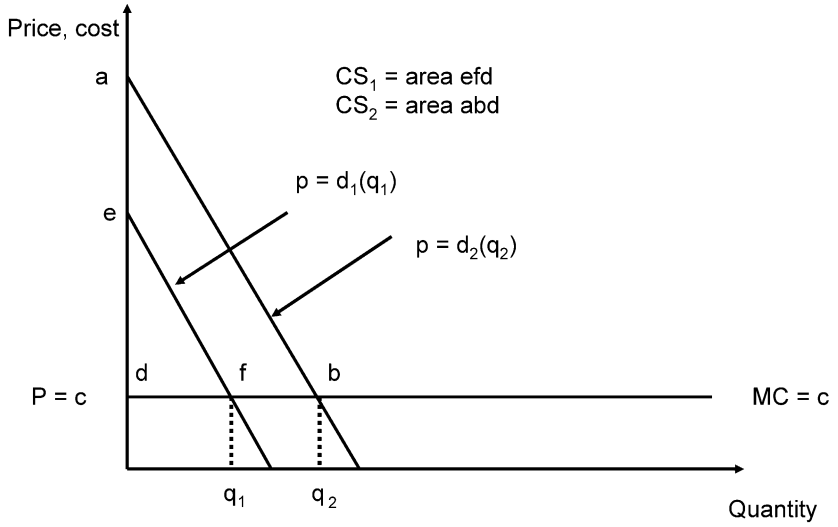


Figure 6. Heterogeneous consumers.

the net surplus for a low-type consumer is $\mathrm{CS}_1 = (S_1 - cq_1)$ and the net surplus for a high-type consumer is $\mathrm{CS}_2 = (S_2 - cq_2)$ where $\mathrm{CS}_1 < \mathrm{CS}_2$ and $n_1\mathrm{CS}_1 + n_2\mathrm{CS}_2 > f_o$. If the regulator were restricted to only a uniform two-part tariff, the highest access charge that could be assessed without forcing the low-value types off of the network would be $A = \mathrm{CS}_1$. If the total revenues generated when all consumers are charged an access fee equal to CS_1 is less than f_o then the break-even constraint would not be satisfied and the product would not be supplied even if it's total value is greater than its total cost. How can we extract more of the consumer's surplus out of the high demand types to cover the regulated firm's budget constraint with the minimum distortion to consumption decisions of both consumer types?

This is a simple example of the more general non-linear pricing problem. Intuitively, we can think of offering a menu of two-part tariffs of the form:

$$T_1 = A_1 + p_1 q_1$$
$$T_2 = A_2 + p_2 q_2$$

Where $A_1 < A_2$ and $p_1 > p_2 \geq c$ as in Figure 7 so that the low demand consumers find it most economical to choose T_1 and the high demand types choose T_2. In order to achieve this *incentive compatibility* property, the tariff T_1 with the low access fee must have a price p_1 that is sufficiently greater than p_2 to make T_1 unattractive to the high demand type. At the same time we would like to keep p_1 and p_2 as close to c as we can to minimize the distortion in consumption arising from prices being greater than marginal cost. The low-demand and high-demand types face a different price on the margin and the optimal prices are chosen to meet the break-even constraint

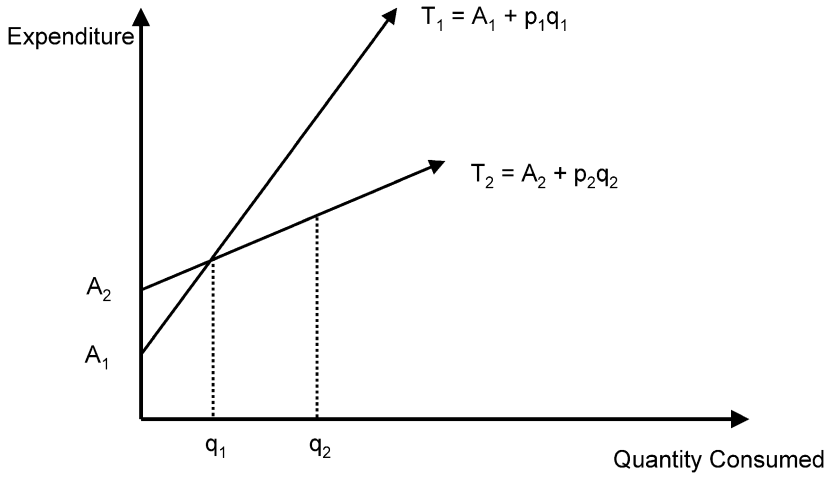


Figure 7. Two-part tariff.

with the minimum distortion. Note, that the menu above is equivalent to a single price schedule that has a single fixed fee A^* and then a usage fee that declines as consumption increases:

$$T(q) = A^* + p_1 q_1 + p_2\left(q_2 - q_1^*\right)$$

for q_1 between 0 and q_1^* and $q_2 > q_1^*$.

Let us turn to a more general case. Following Laffont and Tirole (2000, pp. 70–71) assume that the regulated firm's cost function is as before:

$$C = f_0 + cq$$

There is then a continuum of consumers with different demands for the regulated service and the consumer types are indexed by the parameter θ. A consumer of type θ will be confronted with a non-linear tariff $T(q)$ which has the property that the average expenditure per unit purchased on the service declines as q increases. Assume that the consumer of type θ consumes $q(\theta)$ when she faces $T(q)$ and has net utility $U(\theta) = \theta V[q(\theta)] - T[q(\theta)]$. (Note, this effectively assumes that the distribution of consumer demands shifts outward as θ increases and that the associated individual consumer demand curves do not cross. See Braeutigam (1989) and Brown and Sibley (1986) for more general treatments.)

Assume next that the parameter θ is distributed according to the c.d.f. $G(\theta)$ with density $g(\theta)$ with lower and upper bounds on θ of θ_L and θ_H respectively with the hazard rate $g(\theta)/[1 - G(\theta)]$ increasing with θ. Let $(1 + \lambda)$ denote the shadow cost of the firm's budget constraint. Then maximizing social welfare (gross consumers' surplus

net of the total costs of production) is equivalent to maximizing:

$$\int_{\theta_L}^{\theta_H} \{\theta V[q(\theta)] - T[q(\theta)]\}\, dG(\theta) - (1+\lambda)\int_{\theta_L}^{\theta_H} \{cq(\theta) + k_0 - T[q(\theta)]\}\, dG(\theta) \tag{7}$$

Let $U(\theta) \equiv \theta V[q(\theta)] - T[q(\theta)]$ and we obtain the constrained maximization problem for deriving the properties of the optimal non-linear prices

$$\max \int_{\theta_L}^{\theta_H} \big((1+\lambda)\{\theta V[q(\theta)] - cq(\theta) - k_0\} - \lambda U(\theta)\big)\, dG(\theta) \tag{8}$$

subject to:

$$\dot{U} = V[q(\theta)] \quad \text{and} \quad \dot{q} \geq 0 \tag{9}$$

$$U(\theta \geq 0 \quad \text{for all } \theta \tag{10}$$

where the first constraint is the incentive compatibility constraint and the second constraint is the constraint that all consumers with positive net surplus participate in the market.

Letting $\theta(q) = p(q) = T'(q)$ denote the marginal price that characterizes the optimal non-linear tariff, we obtain

$$\frac{p(q) - c}{p(q)} = \frac{\lambda}{1+\lambda}\frac{1 - G(\theta)}{\theta g(\theta)} \tag{11}$$

which implies that the optimal two-part tariff has the property that the marginal price falls toward marginal cost as θ increases or, alternatively we move from lower to higher demand types.

Willig (1978) shows that any second-best (Ramsey-Boiteux) uniform price schedule can be dominated from a welfare perspective by a non-linear price schedule. In some sense this should not be surprising. By capturing some infra-marginal surplus to help to cover the regulated firm's fixed costs, marginal prices can be moved closer to marginal cost, reducing the pricing distortion, while still satisfying the firm's budget balance constraint.

In fact, non-linear pricing has been used in the pricing of electricity, gas and telephone service since early in the 20th century [Clark (1911, 1913)]. Early proponents of non-linear pricing such as Samuel Insull viewed these pricing methods as a way to expand demand and lower average costs while meeting a break-even constraint [Hughes (1983, pp. 218–226)]. As is the case for uniform prices, the basic structure, though not the level, of the optimal non-linear prices is identical to the structure that would be chosen by a profit maximizing monopoly with the same information about demand patterns and the ability to restrict resale.

6.4. Peak-load pricing

Many public utility services cannot be stored and the demand for these services may vary widely from hour to hour, day to day and season to season. Because these services cannot be stored, the physical capacity of the network must be expanded sufficiently to meet peak demand. Services like electricity distribution and generation, gas distribution, and telephone networks are very capital intensive and the carrying costs (depreciation, interest on debt, return on equity investment) of the capital invested in this capacity is a relatively large fraction of total cost. For example, the demand for electricity varies widely between day and night, between weekdays and weekends and between days with extreme rather than moderate temperatures. Over the course of a year, the difference in demand between peak and trough may be on the order of a factor of three or more. The demand during the peak hour of a very hot day may be double the demand at night on that same day. Since electricity cannot be stored economically, the generating, transmission and distribution capacity of an electric power system must be sufficient to meet these peak demand days, taking into account equipment outages as well as variations in demand. Traditional telephone and natural gas distribution network have similar attributes.

The "peak load pricing" literature, which has been developed primarily in connection with the pricing of electricity, has focused on the specification of efficient prices and investment levels that take account of the variability of demand, the non-storability of the service, the attributes of alternative types of capital equipment available to supply electricity, equipment outages, and the types of metering equipment this is available and at what cost. There is a very extensive theoretical literature on efficient pricing and investment programs for electric power services that was developed mostly during the period 1950–1980 and primarily by French, British and American economists [Nelson (1964), Steiner (1957), Boiteux (1960), Turvey (1968a, 1968b), Kahn (1970), Crew and Kleinforfer (1976), Dreze (1964), Joskow (1976), Brown and Sibley (1986), Panzar (1976), Carlton (1977)]. This theoretical work was applied extensively to the pricing of electricity in France and in England during the 1950s and 1960s. There is little new of late on this topic and I refer interested readers to the references cited above.

The intuition behind the basic peak load pricing results is quite straightforward. If capacity must be built to meet peak demand then when demand is below the peak there will be surplus capacity available. The long run marginal cost of increasing supply to meet an increment in peak demand includes both the additional capital and operating costs of building and operating an increment of peak capacity. The long run marginal cost of increasing supply to meet an increment of off peak demand reflects only the additional operating costs or short run marginal cost of running more of the surplus capacity to meet the higher demand as long as off-peak demand does not increase to a level greater than the peak capacity on the system. Accordingly, the marginal social cost of increasing supply to meet an increase in peak demand will be much higher than the marginal cost of increasing supply to meet an increment of off-peak demand. Efficient price signals should convey these different marginal costs to consumers. Accordingly,

the peak price should be relatively high, reflecting both marginal operating and capital costs, and the off-peak prices low to reflect only the off-peak marginal costs of operating the surplus capacity more intensively.

The following simple model demonstrates this intuitive result and one of several interesting twists to it.

Let $q_D = q_D(p_D)$ = the demand for electricity during day-time hours

and $q_N = q_D(p_N)$ = the demand for electricity during night-time hours

for any $p_D = p_N$ day-time demand is higher than night-time demand ($q_D(p_D) > q_N(p_D)$). The gross surplus during each period (area under the demand curve) is given by $S(q_i)$ and $\frac{\partial S_i}{\partial q_i} = p_i$.

Assume that the production of electricity is characterized by a simple fixed-proportions technology composed of a unit rental cost C_K for each unit of generating capacity (K) and a marginal operating cost C_E for each unit of electricity produced. We will assume that there are no economies of scale, recognizing that any budget balance constraints can be handled with second-best linear or non-linear prices. Demand in any period must be less than or equal to the amount of capacity installed so that $q_D \leq K$ and $q_N \leq K$.

The optimal prices are then given by solving the following program which maximizes net surplus subject to the constraints that output during each period must be less than or equal to the quantity of capacity that has been installed:

$$L^* = S(q_D) + S(q_N) - C_K K - C_E(q_D + q_N) + \lambda_D(K - q_D) + \lambda_N(K - q_N) \tag{12}$$

where λ_D and λ_N are the shadow prices on capacity The first order conditions are then given by:

$$p_D - C_E - \lambda_D = 0$$
$$p_N - C_E - \lambda_N = 0$$
$$\lambda_D + \lambda_N - C_K = 0$$

with complementary slackness conditions

$$\lambda_D(K - q_D) = 0$$
$$\lambda_N(K - q_N) = 0$$

There are then two interesting cases:

Case 1: Classic peak load pricing results:

$$P_D = C_E + C_k \quad (\lambda_D = C_K) \tag{13}$$
$$P_N = C_E \quad (\lambda_N = 0) \tag{14}$$
$$q_N < q_D \tag{15}$$

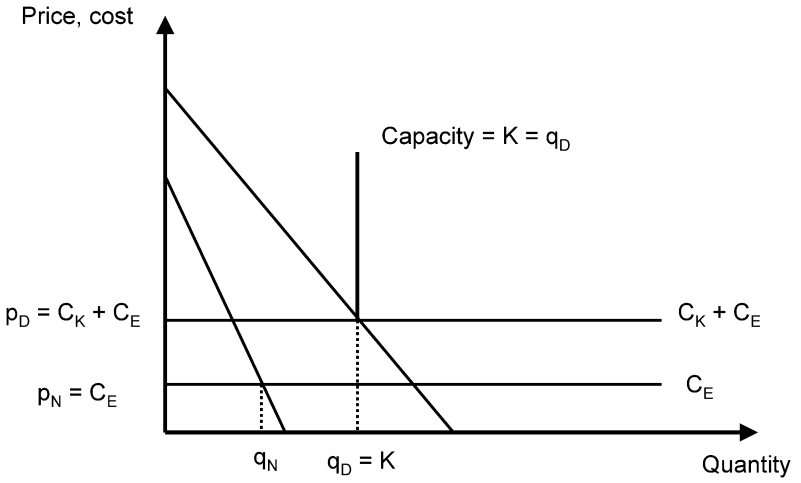


Figure 8. Peak-load pricing.

In this case, the optimal price during the peak period equals the sum of marginal operating costs and marginal capacity costs. In the off-peak period the optimal price equals only marginal operating costs. The result is depicted in Figure 8.

Case 2: Shifting peak case:

$$P_{\mathrm{D}} = C_{\mathrm{E}} + \lambda_{\mathrm{D}} \quad (\lambda_{\mathrm{D}} > 0) \tag{16}$$

$$P_{\mathrm{N}} = C_{\mathrm{E}} + \lambda_{\mathrm{N}} \quad (\lambda_{\mathrm{N}} > 0) \tag{17}$$

$$\lambda_{\mathrm{D}} + \lambda_{\mathrm{N}} - C_{\mathrm{K}} = 0 \tag{18}$$

$$q_{\mathrm{D}} = q_{\mathrm{N}} \tag{19}$$

Here, the optimal prices during the peak and off peak periods effectively share the marginal cost of capacity plus the marginal cost of producing electricity. The peak period price includes a larger share of the marginal cost of capacity than the off-peak price reflecting the differences between consumers' marginal willingness to pay between the two periods. This result is reflected in Figure 9.

The standard case is where $\lambda_{\mathrm{D}} > 0$ and $\lambda_{\mathrm{N}} = 0$. The peak price now equals the marginal capital and operating cost of the equipment and the off-peak price equals only the marginal operating costs. Investment in capacity K is made sufficient to meet peak demand ($\mathrm{K} = q_{\mathrm{D}} > q_{\mathrm{N}}$) and consumers buying power during the peak period pay all of the capital costs. Consumption during the day then carries a higher price than consumption at night. Does this imply that there is price discrimination at work here? The answer is no. Peak and off-peak consumption are essentially separate products and supply in both periods each pay their respective marginal supply costs. What is true, is

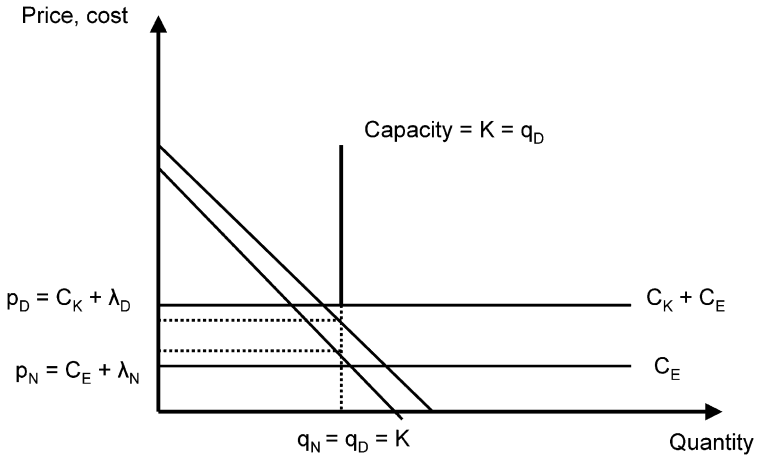


Figure 9. Peak-load pricing with sifting peak.

that the production of peak and off peak supplies are "joint products" that incur joint costs. That is, off-peak supply could not be provided so inexpensively if peak demand was not there to pay all of the capital costs.

The role of joint costs becomes evident when we look at the second potential solution of the simple problem above. This potential solution has the following properties:

$$q_{\mathrm{D}} = q_{\mathrm{N}}$$

and

$$\lambda_{\mathrm{D}} + \lambda_{\mathrm{N}} = C_{\mathrm{K}}$$

In this potential equilibrium, peak and off-peak consumption each bear a share of the capital or capacity costs. This situation arises when the peak and off-peak demands are so elastic that applying the simple peak load pricing rule that the peak demand pays for all capital costs and the off-peak demand for none, ends up shifting the peak demand to the off-peak (night) period. This problem was realized in practice in a number of countries that instituted simple peak load pricing models for electricity during the 1960s and 1970s. The off-peak price was set so low that many consumers installed electric heating equipment that on cold winter nights led electricity demand to reach its peak levels. The solution to the problem is to share the joint capacity costs in a way that reflects the relative valuations of peak and off-peak consumption (at the equilibrium prices) as displayed in Figure 9. The off-peak price still is lower than the peak price, but now all consumption pays a share of the network's capacity costs. Note as well, that the implementation of efficient prices now requires the regulator to have information about the elasticity of demand in different time periods.

There are numerous realistic complications that can and have been introduced into simple peak load pricing models such as the one above. Suppliers can choose among dif-

ferent production techniques with different (in the case of electricity) capital/fuel ratios. In addition, demand cannot be divided simply between "peak" and "off-peak." Rather the system is characterized by continuously varying demands that lie between some lower and upper bound. In this case, since some of the capacity is utilized relatively few hours each year, some during all hours and some for say half the hours of the year, it is economical to install a mix of "base load," "intermediate" and "peak load" capacity [Turvey (1968b), Crew and Kleinforfer (1976), Joskow (1976, 2006)] and by allowing prices to vary with marginal production costs, produce infra-marginal quasi rents to cover some of the costs of investments in production facilities. In addition, it turns out that even ignoring the "shifting peak" issue discussed above, when consumption is priced at marginal operating cost during most time periods, consumers during this hours make a contribution to the capital costs of the network because the marginal operating costs of the now diverse electricity production technologies on the network increases as the demand on the network increases. Enhancements of these models have also considered the stochastic attributes of demand and equipment (unplanned outages and planned maintenance requirements) to derive both optimal levels of reserve capacity and the associated optimal prices with and without real time price variations [Carlton (1977)].

The marginal cost of producing electricity varies almost continuously in real time. And when short run capacity constraints are reached the social marginal cost can jump to the valuation of the marginal consumer who is not served [the value of lost load or the value of unserved energy—Joskow and Tirole (2006)]. Many economists argue that electricity prices should vary in real time to convey better price signals (Borenstein, 2005). However, any judgment about which consumers should pay real time prices must take into account the transactions costs associated with recording consumption in real time, collecting and analyzing the associated data. It is generally thought that for larger customers the welfare gains from better pricing exceed the costs of installing and utilizing more sophisticated meters.

Variable demand, diverse technologies, reliability and real time pricing can all be integrated with the "budget balance" constraint considerations discussed earlier. The same basic second-best pricing results hold, though the relevant marginal costs are now more complicated as is the implementation of the budget balance constraint since when stochastic demand and supply attributes are introduced, the firm's revenues, costs and profits also become uncertain [Crew and Kleinforfer (1986)].

7. Cost of service regulation: response to limited information

The discussion of optimal pricing for a natural monopoly in the last section assumes that the regulator knows all there is to know about the regulated firm's costs and demand. In addition, the regulated firm does not act strategically by changing its managerial effort to increase costs or to distort the information the regulator possesses about its cost opportunities and the demand it faces for the services it provides in response to the incentives created by the regulatory mechanisms that have been chosen. In reality, regulators are not inherently well informed about the attributes of the firm's cost

opportunities, its demand, or its management's effort and performance. The regulated firm knows much more about these variables than does the regulator and, if given the opportunity, may have incentives to act strategically. The firm may provide incorrect information about the its cost, demand and managerial effort attributes to the regulator or the firm may respond to poorly designed regulatory incentives by reducing managerial effort, increasing costs, or reducing the quality of service. Much of the evolution of regulatory agencies and regulatory procedures in the U.S. during the last hundred years has focused on making it possible for the regulator to obtain better information about these variables and to use this information more effectively in the regulatory process. More recent theoretical and empirical research has focused on the development of more efficient regulatory mechanisms that reflect these information asymmetries and associated opportunities for strategic behavior as well as to better exploit opportunities for the regulator to reduce its information disadvantages.

I have chosen to begin the discussion of regulation when the regulator has limited information about the attributes of the firm and its customers with a discussion of traditional "cost of service" or "rate of return regulation" that has been the basic framework for commission regulation in the U.S. during most of the 20th century. The performance of this regulatory process (real and imagined) is the "benchmark" against which alternative mechanisms are compared. The "traditional" cost of service regulation model is frequently criticized as being very inefficient but the way it works in practice is also poorly understood by many of its critics. Its application in fact reflects efforts to respond to imperfect and asymmetric information problems that all regulatory processes must confront. Moreover, the application of modern "incentive regulation" mechanisms is frequently an addition to rather than a replacement for cost-of-service regulation [Joskow (2006)]. After outlining the attributes of cost of service regulation in practice I proceed to discuss the "Averch-Johnson model," first articulated in 1962 and developed extensively in the 1970s and 1980s, which endeavors to examine theoretically the efficiency implications of rate of return regulation and variations in its application.

7.1. Cost-of-service or rate-of-return regulation in practice

U.S. regulatory processes have approached the challenges created by asymmetric information in a number of ways. First, regulators have adopted a *uniform system of accounts* for each regulated industry. These cost-reporting protocols require regulated firms to report their capital and operating costs according to specific accounting rules regarding the valuation of capital assets, depreciation schedules, the treatment of taxes, operating cost categories, allocation of costs between lines of business and between regulated and unregulated activities, and the financial instruments and their costs used by the firm to finance capital investments. These reports are audited and false reports can lead to significant sanctions. Since most U.S. states and federal regulatory agencies use the same uniform system of accounts for firms in a particular industry, the opportunity to perform comparative analyses of firm costs and to apply "yardstick regulation" concepts also becomes a distinct possibility (more on this below). Regulatory agencies also have broad

power to seek additional information from regulated firms that would not normally be included in the annual reports under the uniform system of accounts; for example data on equipment outage and other performance indicia, customer outages, consumer demand patterns, etc., and to perform special studies such as demand forecasting and demand elasticity measurement. These data collection and analysis requirements are one way that U.S. regulators can seek to increase the quality of the information they have about the firms they regulate and reduce the asymmetry of information between the regulator and the firms that it regulates. Whether they use these data and authorities wisely is another matter.

Regulators in the U.S. and other countries have long known, however, that better data and analysis cannot fully resolve the asymmetric information problem. There are inherent differences between firms in terms of their cost opportunities and the managerial skill and effort extended by their managements and traditional accounting methods for measuring capital costs in particular may create more confusion than light. Accordingly, the regulatory process does not require regulators to accept the firms' reported and audited accounting costs as "just and reasonable" when they set prices. They can "disallow" costs that they determine are unreasonably through, for example, independent assessments of firm behavior and comparisons with other comparable firms [Joskow and Schmalensee (1986)]. Moreover, contrary to popular misconceptions, regulated prices are not adjusted to assure that revenues and costs are exactly in balance continuously. There are sometimes lengthy periods of "regulatory lag" during which prices are fixed or adjust only partially in response to realized costs the regulated firm shares in the benefits and burdens of unit cost increases or decreases [Joskow (1973, 1974), Joskow and Schmalensee (1986)]. And regulators may specify simple "incentive regulation" mechanisms that share variations in profitability between the regulated firm and its customers. These are generally called "sliding scale" regulatory mechanisms [Lyon (1996)], a topic that will be explored presently.

The "fixed point" of traditional U.S. regulatory practice is the *rate case* [Phillips (1993)]. The rate case is a public quasi judicial proceeding in which a regulated firm's prices or "tariffs" may be adjusted by the regulatory agency. Once a new set of prices and price adjustment formulas are agreed to with the regulator (and sustains any court challenges) they remain in force until they are adjusted through a subsequent rate case. Contrary to popular characterizations, regulated prices are not adjusted continuously as cost and demand conditions change, but rather are "tested" from time to time through the regulatory prices. Rate cases do not proceed on a fixed schedule but are triggered by requests from the regulated firm, regulators, or third-parties for an examination of the level or structure of prevailing tariffs [Joskow (1973)]. Accordingly, base prices are fixed until they are adjusted by the regulator through a process initiated by the regulated firm or by the regulator on its own initiative (perhaps responding to complaints from interested parties [Joskow (1972)]. This "regulatory lag" between rate cases may be quite long and has implications for the incentives regulated firms have to control costs (Joskow, 1974) and the distribution of surplus between the regulated firm and consumers.

A typical rate case in the U.S. has two phases. The first phase determines the firm's total *revenue requirement* or its total *cost of service*. It is convenient to think of the revenue requirements or cost of service as the firm's budget constraint. The second phase is the *rate design* or *tariff structure* phase. In this phase the actual prices that will be charged for different quantities consumed or to different types of consumers or for different products is determined. It is in the rate design phase where concepts of Ramsey pricing, non-linear pricing and peak load pricing would be applied in practice.

The firm's revenue requirements or cost of service has numerous individual components which can be grouped into a few major categories:

a. Operating costs (e.g. fuel, labor, materials and supplies)—OC
b. Capital related costs that define the effective "rental price" for capital that will be included in the firm's total "cost of service" for any given time period. These capital related charges are a function of:
   i. the value of the firm's "regulatory asset base" or its "rate base" (RAV)
   ii. the annual amount of depreciation on the regulatory asset base (D)
   iii. the allowed rate of return (s) on the regulatory asset base
   iv. income tax rate (t) on the firm's gross profits
c. Other costs (e.g. property taxes, franchise fees)—F

7.1.1. Regulated revenue requirement or total cost of service

The regulated firm's total *revenue requirements* or *cost of service* in year t is then given by

$$R_t = \mathrm{OC}_t + D_t + r(1+t)\mathrm{RAV} + F_t \tag{20}$$

These cost components are initially drawn from the regulated firm's books and records based on a uniform system of accounts adopted by the regulator. An important part of the formal rate case is to evaluate whether the firm's costs as reported on its books or projected into the future are "reasonable." The regulatory agency may rely on its own staff's evaluations to identify costs that were "unreasonable" or unrepresentative of a typical year, or the regulator may also rely on studies presented by third-party "intervenors" in the rate case [Joskow (1972)]. Interested third-parties are permitted to participate fully in a rate case and representatives of different types of consumers, a public advocate, and non-government public interest groups often participate in these cases, as well as in any settlement negotiations that are increasingly relied upon to cut the administrative process short. Costs that the regulatory agency determines are unreasonable are then "disallowed" and deducted from the regulated firm's cost of service.

There are a number of methods available to assess the "reasonableness" of a firm's expenditures. One type of approach that is sometimes used is a "yardstick" approach in which a particular firm's costs are compared to the costs of comparable firms and significant deviations subject to some disallowance [e.g. Jamasb and Pollitt (2001, 2003), Carrington, Coelli, and Groom (2002), Schleifer (1985)]. Such an approach has been

used to evaluate fuel costs, labor productivity, wages, executive compensation, construction costs and other costs. A related approach is to retain outside experts to review the firm's expenditure experience in specific areas and to opine on whether they were reasonably efficient given industry norms. The regulator may question assumptions about future demand growth, the timing of replacements of capital equipment, wage growth, etc. Finally, accountants comb through the regulated firm's books to search for expenditures that are either prohibited (e.g. Red Sox tickets for the CEO's family) or that may be of questionable value to the regulated firm's customers (e.g. a large fleet of corporate jets). These reasonableness review processes historically tended to be rather arbitrary and ad hoc in practice, but have become more scientific over time as benchmarking methods have been developed and applied. Since the regulated firm always knows more about its own cost opportunities and the reasons why it made certain expenditures than does the regulator, this process highlights the importance of thinking about regulation from an asymmetric information perspective.

From the earliest days of rate or return regulation, a major issue that has been addressed by academics, regulators and the courts is the proper way to value the firm's assets in which it has invested capital and how the associated depreciation rates and allowed rate of return on investment should be determined [Sharfman (1928), Phillips (1993), Bonbright (1961), Clemens (1950)]. A regulated firm makes investments in long-lived capital facilities. Regulators must determine how consumers will pay for the costs of these facilities over their economic lives. A stock (the value of capital investments) must be transformed into a stream of cash flows or annual rental charges over the life of the assets in which the regulated firm has invested in order to set the prices that the firm can charge and the associated revenues that it will realize to meet the firm's overall budget constraint.

The basic legal principle that governs price regulation in the U.S. is that regulated prices must be set at levels that give the regulated firm a reasonable opportunity to recover the costs of the investments it makes efficiently to meet its service obligations and no more than is necessary to do so.[15] One way of operationalizing this legal principle is to reduce it to the rule that the present discounted value of future cash flows that flow back to investors in the firm (equity, debt, preferred stock) should be at least equal to the cost of the capital facilities in which the firm has invested,. Where the discount rate is the firm's risk adjusted cost of capital "r." Let:

Π_t = cash flow in year t = Revenues − operating costs − taxes and other expenses

K_0 = the "reasonable" cost of an asset added by the utility in year t

r = the firm's after tax opportunity cost of capital

[15] *Federal Power Commission v. Hope Natural Gas Co., 320 U.S. 591, 602 (1944).*

The basic rule for setting prices to provide an appropriate return of and on an investment in an asset with a cost of K_0 made by the regulated firm can then be defined as:

$$K_0 \leq \sum_{1}^{n} \frac{\pi_t}{(1+r)^t} \tag{21}$$

where n is the useful/economic life of the asset. If this condition holds, then the regulated firm should be willing to make the investment since it will cover its costs, including a return on its investment greater than or equal to its opportunity cost of capital. If the relationship holds with equality then consumers are asked to pay no more than is necessary to attract investments in assets required to provide services efficiently. Since a regulated firm will typically be composed of many assets reflecting investments in capital facilities made at many different times in the past, this relationship must hold both on the margin and in the aggregate for all assets. However, it is easiest to address the relevant issues by considering a single-asset firm, say a natural gas pipeline company, with a single productive asset that works perfectly for n years and then no longer works at all (a one-horse-shay model).

There are many (infinite) different streams of cash flows that satisfy the NPV condition in (21) for a single asset firm that invests in the asset in year 1 and uses it productively until it is retired in year n. These cash flows can have many different time profiles. Cash flows could start high and decline over time. Cash flows could start low and rise over time. Cash flows could be constant over the life of the asset. Much of the historical debate about the valuation of the regulatory asset base, the depreciation rate and rate of return values to be used to turn the value of the capital invested by the regulated monopoly firm in productive assets into an appropriate stream of cash flows over time, has reflected alternative views about the appropriate time profile and (perhaps unintended) the ultimate level of the present discounted value of these cash flows. Unfortunately, this debate about asset valuation, depreciation and allowed return was long on rhetoric and short on mathematical analysis, had difficulty dealing with inflation in general and confused real and nominal interest rates in particular [Sharfman (1928), Clemens (1950, Chapter 7)]. The discussion and resulting regulatory accounting rules also assume that the regulated firm is a monopoly, does not face competition, and will be in a position to charge prices that recover the cost of the investment over the accounting life of the asset. Giving customers the option to switch back and forth from the regulated firms effectively imbeds a costly option into the "regulatory contract" and requires alternative formulas for calculating prices that yield revenues with an expected value equal to the cost of the investment [Pindyck (2004), Hausman (1997), Hausman and Myers (2002)].

A natural starting point for an economist is to rely on economic principles to value the regulated firm's assets. We would try to calculate a pattern of rental prices for the firm's assets that simulates the trajectory of rental prices that would emerge if the associated capital services were sold in a competitive market. This approach implies valuing assets at their competitive market values, using economic depreciation concepts, and taking

appropriate account of inflation in the calculation of real and nominal interest rates. Consider the following simple example:

Assume that a machine producing a homogeneous product depreciates (physically) at a rate d per period. You can think of this as the number of units of output from the machine declining at a rate d over time. Assume that operating costs are zero. Define the competitive rental value for a new machine at any time s by $v(s)$. Then in year s the rental value on an old machine bought in a previous year t would be

$$v(s)\mathrm{e}^{-d(s-t)}$$

since $(s-t)$ is the number of years the machine has been decaying.

The price of a new machine purchased in year t $[P(t)]$ is the present discounted value of future rental values. Let r be the firm's discount rate (cost of capital). Then the present value of the rental income in year s discounted back to year t is

$$\mathrm{e}^{-r(s-t)}v(s)\mathrm{e}^{-d(s-t)} = \mathrm{e}^{(r+d)t}v(s)\mathrm{e}^{-(r+d)s}$$

and the present value of the machine in year t is:

$$\begin{aligned}\mathrm{PDV}(t) &= \int_t^\infty \mathrm{e}^{(r+d)t}v(s)\mathrm{e}^{-(r+d)s}\,\mathrm{d}s \\ &= \text{competitive market price for a new machine in year } t \\ &= P(t) \qquad (22)\end{aligned}$$

Rewrite this equation as:

$$P(t) = \mathrm{e}^{(r+d)t}\int_t^\infty v(s)\mathrm{e}^{-(r+d)s}\,\mathrm{d}s \qquad (23)$$

and differentiate with respect to t

$$\mathrm{d}P(t)/\mathrm{d}t = (r+d)P(t) - v(t)$$

or

$$v(t) = (r+d)P(t) - \mathrm{d}P(t)/\mathrm{d}t$$

where $\mathrm{d}P(t)/\mathrm{d}t$ reflects exogenous changes in the price of new machines over time. These price changes reflect general inflation (i) and technological change (δ) leading to lower cost machines (or more productive machines). The changes in the prices of new machines affect the value of old machines because new machines must compete with old machines producing the same product.

$$\begin{aligned}v(t) &= (r+d)P(t) - (i-\delta)P(t) \\ &= (r+d-i+\delta)P(t) \qquad (24)\end{aligned}$$

The economic depreciation rate is then $(d-i+\delta)$ and the allowed rate of return consistent with it is given by r the firm's nominal cost of capital. Both are applied to the current competitive market value of the asset $P(t)$.

Equation (24) provides the basic formula for setting both the level and time profile of the capital cost component of user prices for this single-asset regulated firm assuming that there is a credible regulatory commitment to compensate the firm in this way over the entire economic life of the asset. Even though the value of the regulatory asset is effectively marked to market on a continuing basis, the combination of sunk costs and asset specificity considerations would require a different pricing arrangement if, for example, customers were free to turn to competing suppliers if changing supply and demand conditions made it economical to do so [Pindyck (2004), Hausman (1997), Hausman and Myers (2002)].

The earliest efforts to develop capital valuation and pricing principles indeed focused on "fair value" rate base approaches in which the regulated firm's assets would be revalued each year based on the consideration of "reproduction cost," and other "fair market value" methods, including giving some consideration to "original cost" [Troxel (1947, Chapters 12 and 13), Clemens (1950, Chapter 7), Kahn (1970, pp. 35–45)]. Implementing these concepts in practice turned out to be very difficult with rapid technological change and widely varying rates of inflation. Regulated firms liked "reproduction cost new" methods for valuing assets when there was robust inflation (as during the 1920s), but not when the nominal prices of equipment were falling (as in the 1930s). Moreover, "fair market value" rules led regulated firms to engage in "daisy chains" in which they would trade assets back and forth at inflated prices and then seek to increase the value of their rate bases accordingly. Methods to measure a firm's cost of capital were poorly developed [Troxel (1947, Chapters 17, 18, 19)]. Many regulated firm asset valuation cases were litigated in court. The guidance given by the courts was not what one could call crystal clear [Troxel (1947, Chapter 12)].

Beginning in the early 1920s, alternatives to the "fair value" concept began to be promoted. In a dissenting opinion in 1923,[16] Justice Louis Brandeis criticized the "fair valuation" approach. He proposed instead a formula that is based on what he called the prudent investment standard. Regulators would first determine whether an investment and its associated costs reflected "prudent" or reasonable decisions by the regulated firm. If they did, investors were to be permitted to earn a return of and on the "original cost" of this investment. The formula for determining the trajectory of capital related charges specified that regulators should use straight line depreciation of the original cost of the investment, value the regulatory asset base at the original cost of plant and equipment prudently incurred less the accumulated depreciation associated with it at any particular point in time, and apply an allowed rate of return equal to the firm's nominal cost of capital.

Consider a single asset firm with a prudent investment cost of K_0 at time zero. The Brandeis formula would choose an accounting life N for the asset. The annual depreciation was then given by $D_t = K_0/N$. The regulatory asset base in any year was then given by $\text{RAV}_t = K_0 - \sum_0^t D_t$. Then prices are set to produce net cash flows (after

[16] *Southwestern Bell Telephone Company v. Public Service Commission of Missouri 262 U.S. 276 (1923).*

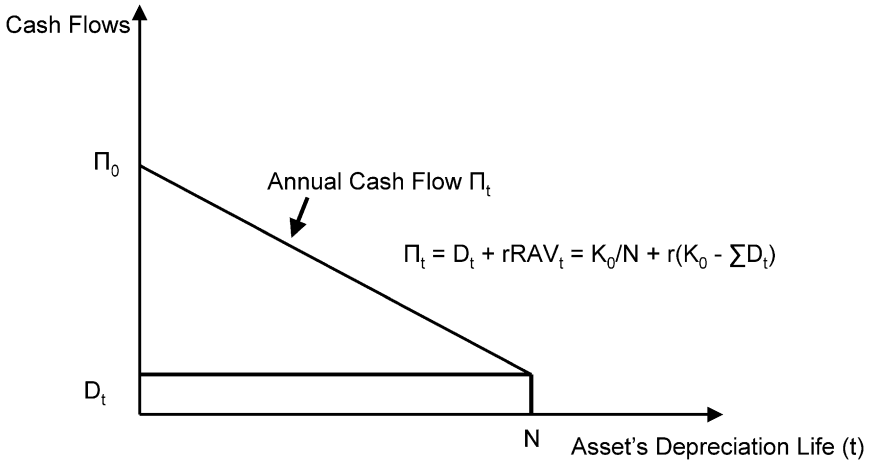


Figure 10. Depreciated original cost rate base.

operating costs, taxes and other allowable fees) based on the following net cash flow formula

$$\Pi_t = D_t + r\mathrm{RAV}_t = K_0/N + r\left(K_0 - \sum_0^t D_t\right) \tag{25}$$

which can be easily extended to multiple assets with different in-service dates and service lives. The cash flow profile for a single-asset firm is displayed in Figure 10. Brandeis argued that this approach would make it possible for regulators and the courts to "avoid the 'delusive' calculations, 'shifting theories," and varying estimates that the engineers use as they measure the reproduction costs and present values of utility properties." [Troxel (1947, p. 271)] while providing regulated firms with a fair return on the prudent cost of investments that they have made to support the provisions of regulated services.

The Brandeis formula is quite straightforward, and the prudent investment standard compatible with a regulatory system that guards against regulatory hold-ups of investors ex post. However, does it satisfy the NPV criterion discussed earlier and, in this way, provide an expected return that is high enough to attract investment, but not so high that it yield prices significantly higher than necessary to attract investment? It turns out that the Brandeis formula satisfies the NPV criterion (Schmalensee, 1989a). The present discounted value of cash flows calculated using the Brandeis formula (including an allowed rate of return that is equal to the regulated firm's nominal opportunity cost of capital) is exactly equal to the original cost of the investment; investors get a return of their investment and a return on their investment equal to their opportunity cost of capital. As Brandeis suggested, it provide a simple and consistent method for compen-

sating investors for capital costs and eliminates the uncertainties and opportunities for manipulation that characterized the earlier application of "fair valuation" concepts.

Beginning in the 1930s, regulators began to adopt and the courts began to accept the prudent investment/original cost approach and by the end of World War II it became the primary method for determining the capital charge component of regulated prices. In the *Hope* decision in 1944, the Supreme Court concluded that from a Constitutional perspective it was the "result" that mattered rather than the choice of a particular method and, in this way, getting the courts disentangled from deciding whether or not specific details of the regulatory formulas chosen by state and federal regulators passed Constitutional muster.

> "Under the statutory standard of 'just and reasonable it is the result reached not the method employed which is controlling."[17]
> "Rates which enable the company to operate successfully, to maintain its financial integrity, to attract capital, and to compensate its investors for the risk assumed certainly cannot be condemned as invalid, even though they might produce only a meager return on the so-called 'fair value' rate base."[18]

While the prudent investment/depreciated original cost standard satisfies the NPV criterion, and may have other attractive properties for attracting investment to regulated industries, it also has some peculiarities. These can be seen most clearly for the single asset company (e.g. a pipeline). The time pattern of capital charges has the property of starting at a particular level defined by the undepreciated (or barely depreciated) RAV equal to a value close to K_0 and then declining continuously over the life of the asset until it approaches zero at the end of its useful life (see Figure 10). However, there is no particular reason to believe that the annual capital charges defined by the formula *at any particular point in time*, reflect the "competitive" capital charges or rental rates that would emerge in a competitive market. For example, if we apply the economic depreciation and competitive market value RAV formula discussed earlier, if there is inflation but no technological change, the capital charges for the asset should increase at the rate of inflation over time rather than decrease steadily as they do with the Brandeis formula. In this case if we use the Brandeis formula, regulated prices start out too high and end up too low when the Brandeis formula is applied in this case. If the asset is replaced in year $N + 1$ and the Brandeis formula applied de novo to the new asset, the price for capital related charges will jump back to the value in year 1 (assuming zero inflation and no technological change) and then gradually decline again over time. Thus, while the Brandeis formula gives the correct NPV of cash flows to allow for recovery of a return of and on investment, it may also yield the wrong prices (rental charges associated with capital costs) at any particular point in time. This in turn can lead to the standard consumption distortions resulting from prices that are too high or too low.

[17] *Federal Power Commission v. Hope Natural Gas Co., 320 U.S. 591, 602 (1944).*
[18] Ibid at 605.

Moreover, because assets do not reflect their market value at any particular point in time, the Brandeis formula can and has led to other problems. Regulated prices for otherwise identical firms may be very different because the ages of their assets happen to be different from one another even if their market values are the same. An old coal-fired power plant may have a much higher market value than a new oil-fired power plant, but the prices charged to consumers of the regulated firm with the old coal plants will be low while those of the utility with the new oil-fired plant may be high. As an asset ages, the capital charges associated with it approach zero. For a single asset company, when this asset is replaced at an original cost reflecting current prices, the application of the Brandeis formula leads to a sudden large price increase (known as "rate shock") which creates both consumption distortions and political problems for regulators. Finally, when assets are carried at values significantly greater than their market values, it may create incentives for inefficient entry as well as transition problems when competition is introduced into formerly regulated industries. Who pays for the undepreciated portion of the new oil plant that has a low competitive market value and who gets the benefits from deregulating the old coal plant whose market value is much higher than its RAV when competition replaces regulated monopoly? These so-called "stranded cost" and "stranded benefit" attributes [Joskow (2000)] of the Brandeis formula have plagued the transitions to competition in telecommunications (e.g. mechanical switches that were depreciated too slowly in the face of rapid technological change) and electric power (e.g. costly nuclear power plants that were "prudent" investments when they were made).

It turns out that any formula for calculating the annual capital or rental charge component of regulated prices that has the property (a) the firm earns its cost of capital each period on a rate base equal to the depreciated original cost of its investments and (b) earns the book depreciation deducted from the rate base in each period, satisfies the NPV and investment attraction properties of the Brandeis formula [Schmalensee (1989a)]. There is nothing special about the Brandeis formula in this regard. Alternative formulas that have capital charges for an asset rise, fall or remain constant over time can be specified with the same NPV property. So, in principle, the Brandeis formula could be adjusted to take account of physical depreciation, technological change and inflation to better match both the capital attraction goals and the efficient pricing goals of good regulation.

Note that if a capital investment amortization formula of this type is used, the present discounted value of the firm's net cash flows using the firm's cost of capital as the discount rate should equal its regulatory asset base (RAV) or what is often referred to as its regulatory "book value" B. If investors value the firm based on the net present value of its expected future cash flows using the firm's discount rate then the regulated firm's market value M should be equal to its book value B at any point in time. Accordingly, a simple empirical test is available to determine whether the regulatory process is expected by investors to yield returns that are greater than, less than or equal to the firm's cost of capital. This test involves calculating the ratio of the firm's market value to its

book value:

$M/B = 1 \rightarrow$ Expected returns equal to the cost of capital

$M/B > 1 \rightarrow$ Expected returns greater than the cost ofcapital

$M/B < 1 \rightarrow$ Expected returns less than the cost of capital

In the presence of regulatory lag, we would not expect the M/B always to be equal to one. Moreover, as we shall discuss presently, there may be very good incentive reasons to adopt incentive regulatory mechanisms that, on average, will yield returns that exceed a typical firm's cost of capital. In fact, M/B ratios for regulated electric utilities have varied widely over time [Joskow (1989), Greene and Smiley (1984)] though during most periods of time they have exceeded 1. This is consistent with the observation I made earlier. Due to regulatory lag, a regulated firm's prices are not adjusted continuously to equal its actual costs of production. Deviations between prices and costs may persist for long periods of time [Joskow (1972, 1974)] and have significant effects on the regulated firm's market value [Joskow (1989)]. Accordingly, regulatory lag has both incentive effects and rent extraction effects that are often ignored in uniformed discussions of traditional cost of service regulation.

The final component of the computation of the capital charges that are to be included in regulated prices involves the calculation of the allowed rate of return on investment. Regulatory practice is to set a "fair rate of return" that reflects the firm's nominal cost of capital. Regulated firms are typically financed with a combination of debt, equity and preferred stock [Spiegel and Spulber (1994), Myers (1972a, 1972b)]. The allowed rate of return is typically calculated as the weighted average of the interest rate on debt, preferred stock and an estimate of the firm's opportunity cost of capital, taking the tax treatment of interest payments and the taxability of net income that flows to equity investors. So consider a regulated firm with the following capital structure:

Instrument	Average coupon rate	Fraction of capitalization
Debt	8.0%	50%
Preferred stock	6.0%	10%
Equity	N/A	40%

Then the firm's weighted average cost of capital (net of taxes) is given by

$$r = 8.0 * 0.5 + 6.0 * 0.1 + r_e * 0.4 \tag{26}$$

where r_e is the firm's opportunity cost of equity capital which must then be estimated. Rate cases focus primarily on estimating the firm's opportunity cost of equity capital and, to a lesser degree, determining the appropriate mix of debt, preferred stock, and equity that composes the firm's capital structure. A variety of methods have been employed to measure the regulated firm's cost of equity capital [Myers (1972a, 1972b)], including the so-called discounted cash flow model, the capital asset pricing model, and other "risk premium" approaches [Phillips (1993, pp. 383–412)]. At least in the U.S.,

the methods that are typically used to estimate the regulated firm's cost of capital are surprisingly unsophisticated given the advances that have been made in theoretical and empirical finance in the last thirty years.

With all of these cost components computed the regulator adds them together to determine the firm's "revenue requirement" or total "cost of service" R. This is effectively the budget balance constraint used by the regulator to establish the level and structure of prices—the firm's "tariff"—for the services sold by the regulated firm.

7.1.2. Rate design or tariff structure

In establishing the firm's tariff structure or rate design, regulators typically identify different "classes" of customers, e.g. residential, commercial, farm, and industrial, which may be further divided into further sub-classes (small commercial, large commercial, voltage level differentiation) [Phillips (1993, Chapter 10)]. Since regulatory statutes often require that prices not be "unduly discriminatory," the definition of tariff classes typically is justified by differences in the costs of serving the different groups. In reality, the arbitrariness of allocating joint costs among different groups of customers provides significant flexibility for regulators to take non-cost factors into account [Bonbright (1961), Clemens (1950, Chapter 11), Salinger (1998)]. For example, residential electricity customers require a costly low-voltage distribution system while large industrial customers take power from the network at higher voltages and install their own equipment to step down the voltage for use in their facilities. Accordingly, since the services provided to residential customers have different costs from those provided to large industrial customers it makes economic sense to charge residential and industrial customers different prices. At the same time, large industrial customers may have competitive alternatives (e.g. self-generation of electricity, shifting production to another state with lower prices) that residential customers do not have and this sets the stage for third-degree price discrimination. Many states have special rates for low-income consumers and may have special tariffs for particular groups of customers (e.g. steel mills), reflecting income distribution and political economy considerations. Historically, income redistribution and political economy considerations played a very important role in the specification of telephone services. Local rates were generally set low and long distance rates high, just the opposite of what theories of optimal pricing would suggest [Hausman, Tardiff, and Belinfante (1993), Crandall and Hausman (2000)] and prices in rural areas where set low relative to the cost of serving these customers compared to the price cost margins in urban areas. The joint costs associated with providing both local and long distance services using the same local telephone network made it relatively easy for federal and state regulators to use arbitrary allocations of these costs to "cost justify" almost any tariff structure that they thought met a variety of redistributional and interest group politics driven goals (Salinger, 1998). Non-linear prices have been a component of regulated tariffs for electricity, gas and telephone services since these services first became available. What is clear, however, is that the formal application of the theoretical principles behind Ramsey-Boiteux pricing, non-linear pricing, and peak-load pricing has been used infre-

quently by U.S. regulators, while these concepts have been used extensively in France since the 1950s [Nelson (1964)].

7.2. The Averch-Johnson model

What has come to be known as the Averch-Johnson or "A-J" model [Averch and Johnson (1962), Baumol and Klevorick (1970), Bailey (1973)] represents an early effort to capture analytically the potential effects of rate of return regulation on the behavior of a regulated monopoly. The A-J model begins with a profit-maximizing monopoly firm with a neoclassical production function $q = F(K, L)$ and facing an inverse market demand curve $p = D(q)$. The firm invests in capital (K) with an opportunity cost of capital r (the price of capital is normalized to unity and there is no depreciation) and hires labor L at a wage w. The monopoly's profits are given by:

$$\Pi = D(q)q - wL - rK$$

It is convenient to write the firm's revenues in terms of the inputs K and L that are utilized to produce output q. Let the firms total revenue $R = R(K, L)$ and then

$$\Pi = R(K, L) - wL - rK \tag{27}$$

The regulator has one instrument at it's disposal to control the monopoly's prices. It can set the firm's "allowed rate of return" on capital s at a level greater than or equal to the firm's opportunity cost of capital r and less than the rate of return r_m that would be earned by an unregulated monopoly. The firm's variable costs wL and its capital charges sK are passed through into prices continuously and automatically without any further regulatory review or delay. The regulator has no particular objective function and is assumed only to know the firm's cost of capital r. It has no other information about the firm's production function, it's costs or its demand. The rate of return constraint applied to the firm is then given by:

$$\big[R(K, L) - wL - sK\big] \leq 0 \quad \text{where } r < s < r_m$$

or rewriting

$$\Pi \leq (s - r)K \tag{28}$$

The regulated firm is then assumed to maximizes profits (1) subject to this rate of return constraint (2). Assuming that the rate of return constraint is binding and that a solution with $q > 0$, $K > 0$ and $L > 0$ exists, the firm's constrained maximization problem becomes:

$$\underset{(K,L,\lambda)}{\text{Max}}\, \Pi^* = R(K, L) - wL - rK - \lambda\big[R(K, L) - wL - sK\big] \tag{29}$$

where λ is the shadow price of the constraint. The first order conditions are

$$\frac{\partial \Pi^*}{\partial K} = R_K - r - \lambda(R_K - s) = 0 \tag{30}$$

$$\frac{\partial \Pi^*}{\partial L} = R_L - w - \lambda(R_L - w) = 0 \tag{31}$$

$$\frac{\partial \Pi^*}{\partial \lambda} = R(K, L) - wL - sK = 0 \tag{32}$$

where R_K and R_L are the marginal revenue products of capital and labor respectively ($R_i = \mathrm{MR}_q F_i$). We can rewrite these conditions as

$$R_K = r + [\lambda/(1+\lambda)](r - s)$$

$$R_L = w$$

and (using the second order conditions) $0 < \lambda < 1$. From the first order conditions we can derive the regulated firm's marginal rate of technical substitution of capital for labor as

$$\mathrm{MRT}_{KL} = F_K/F_L = r/w + \lambda/(1+\lambda)\big[(r-s)/w\big] \tag{33}$$

This leads to the primary A-J results. A cost minimizing firm would equate the marginal rate of substitution of capital for labor to the input price ratio. Accordingly, the regulated monopoly operating subject to a rate of return constraint does not minimize costs—input proportions are distorted. Indeed with $0 < \lambda < 1$, the distortion is in a particular direction. Since $\mathrm{MRT}_{KL} < r/w$ for the equilibrium level of output the regulated firm uses too much capital relative to labor. This is sometimes referred to as the *capital using bias* of rate of return regulation. Basically, the rate of return constraint drives a wedge between the firm's actual cost of capital and its effective net cost of capital after taking account of the net benefits associated with increasing the amount of capital used when there is a net return of $(s - r)$ on the margin from adding capital, other things equal.

During the 1970s, many variations on the original A-J model appeared in the literature to extend these results. The reader is referred to Baumol and Klevorick (1970), Klevorick (1973) and Bailey (1973) for a number of these extension. Among the additional results of interest are:

(a) The A-J firm does not "waste" inputs in the sense that inputs are hired but are not put to productive use [Bailey (1973)]. The firm produces on the boundary of its production function and there is no "X-inefficiency" or waste in that sense. The inefficiency is entirely in terms of inefficient input proportions.

(b) As the allowed rate of return s approaches the cost of capital r, the magnitude of the input distortion increases [Baumol and Klevorick (1970)].

(c) There is an optimal value s* for the allowed rate of return that balances the benefits of lower prices against the increased input distortions from a lower allowed rate of return [Klevorick (1971), Sheshinski (1971)]. However, to calculate the optimal rate of return the regulator would have to know the attributes of the firm's production function, input prices and demand, information that the regulator is assumed not to possess. If the regulator did have this information she could simply calculate the optimal input proportions and penalize the firm from deviating from them.

(d) Introducing "regulatory lag" into the model (in a somewhat clumsy fashion) reduces the magnitude of the input bias [Baumol and Klevorick (1970), Bailey and Coleman (1971), Klevorick (1973)]. This is the case because if prices are fixed between rate cases, the firm can increase its profits by reducing its costs [Joskow (1974)] until the next rate review when the rate of return constraint would be applied again. If rate of return reviews are few and far between the firm essentially becomes a residual claimant on cost reductions and has powerful incentives to minimize costs. In this case, rate of return regulation has incentive properties similar to "price cap" regulation with "resets" every few years. Price cap regulation will be discussed further below.

(e) Rate of return regulation of this type can affect the profitability of peak load pricing. In particular, under certain conditions peak load pricing may reduce the firm's capital/labor ratio and it could be more profitable for the firm not to level out demand variations. However, the A-J effect could go in the other direction as well.

A lot of ink was spent on the many papers that developed variations on the A-J model and to test its implications empirically during the 1970s and 1980s. The major conceptual innovation of this literature was to highlight the possibility that regulatory mechanisms could create incentives for regulated firms to produce inefficiently and, perhaps, to adopt organization forms (e.g. vertical integration) and pricing strategies (e.g. peak load pricing) that are not optimal. Moreover, these results depend upon an extreme asymmetry of information between the regulated firm and the regulator (or just the opposite if we assume that the regulator can set the optimal s^*). In the A-J type models, the regulator knows essentially nothing about the firm's cost opportunities, realized costs, or demand. It just sets an allowed rate of return and the firm does its thing. Imperfect and asymmetric information are important attributes of regulation from both a normative and a positive perspective. However, implicitly assuming the regulators have no information is an extreme case. Beyond this, there are significant deviations between the model's assumptions (as advanced through the literature) and how regulators actually regulate. Efforts to introduce dynamics and incentive effects through regulatory lag have been cumbersome within this modeling framework. Empirical tests have not been particularly successful [Joskow and Rose (1989)]. Moreover, the particular kind of inefficiency identified by the model (inefficient input proportions) is quite different from the kind of managerial waste and inefficiency that concerns policymakers and has been revealed in the empirical literature on the effects of regulation and privatization—X-inefficiency of various types arising from imperfections in managerial efforts to minimize the costs of production, leading to production inside the production frontier and not just at the wrong location on the production frontier.

8. Incentive regulation: theory

8.1. Introduction

It should be clear by now that regulators face a number of challenges in achieving the public interest goals identified at the end of Section 3. The conventional theories of optimal pricing, production and investment by regulated firms assume that regulators are completely informed about the technology, costs and consumer demand attributes facing the firms they regulate. This is clearly not the case in reality. Regulators have imperfect information about the cost opportunities and behavior of the regulated firm and the attributes of the demand for its services that it faces. Moreover, the regulated firm generally has more information about these attributes than does the regulator or third parties which may have incentives to provide the regulator with additional information (truthful or untruthful) about the regulated firm. Accordingly, the regulated firm may use its information advantage strategically in the regulatory process to increase its profits or to pursue other managerial goals, to the disadvantage of consumers [Owen and Brauetigam (1978)]. These problems may be further exacerbated if the regulated firm can "capture" the regulatory agency and induce it to give more weight to its interests [Posner (1974), McCubbins (1985), Spiller (1990), Laffont and Tirole (1993, Chapter 5)]. Alternatively, other interest groups may be able to "capture" the regulator and, in the presence of long-lived sunk investments, engage in "regulatory holdups" or expropriation of the regulated firm's assets. Higher levels of government, such as the courts and the legislature, also have imperfect information about both the regulator and the regulated firm and can monitor their behavior only imperfectly [McNollgast (Chapter 22 in this handbook)].

The evolution of regulatory practices in the U.S. reflects efforts to mitigate the information disadvantages that regulators must deal with, as well as broader issues of regulatory capture and monitoring by other levels of government and consumers. As already noted, these institutions and practices are reflected in laws and regulations that require firms to adhere to a uniform system of accounts, give regulators access to the books and records of the regulated firm and the right to request additional information on a case by case basis, auditing requirements, staff resources to evaluate this information, transparency requirements such as public hearings and written decisions, ex parte communications rules, opportunities for third parties to participate in regulatory proceedings to (in theory)[19] assist the regulatory agency in developing better information and reducing its information disadvantage, appeals court review, and legislative oversight processes. In addition, since regulation is a repeated game, the regulator (as well as legislators and appeals courts) can learn about the firm's attributes as it observes its behavioral and performance responses to regulatory decisions over time and, as a result,

[19] Of course, third parties may have an incentive to inject inaccurate information into the regulatory process as well.

the regulated firm naturally develops a reputation for the credibility of its claims and the information that it uses to support them. However, although U.S. regulatory practice focused on improving the information available to regulators, the regulatory mechanisms adopted typically did not utilize this information as effectively as they could have until relatively recently.

The A-J model and its progeny are, in a sense, the first crude analytical efforts to understand how, when regulators are poorly informed and have limited instruments at their disposal, the application of particular mechanisms to constrain the prices charged by a regulated firm may create incentives for a firm to respond in ways that lead to inefficiencies in other dimensions; in the AJ-type models to depart from cost-minimizing input proportions with a bias towards using more capital. However, in the A-J model the regulator has essentially no information about the regulated firm's costs or demand, there is no specification of the objectives of and incentives faced by the firm's managers that might lead the firm to exhibit inefficiencies in other dimensions, the instruments available to the regulatory are very limited, and indeed the choice of mechanisms by the regulator does not flow from a clear specification of managerial objectives and constraints.

More recent work on the theory of optimal incentive regulation deals with asymmetric information problems, contracting constraints, regulatory credibility issues, dynamic considerations, regulatory capture, and other issues that regulatory processes have been trying to respond to for decades much more directly and effectively [Laffont and Tirole (1993), Armstrong, Cowan, and Vickers (1994), Armstrong and Sappington (2003a, 2003b)]. This has been accomplished by applying modern theories of the firm, incentive mechanism design theory, auction theory, contract theory, and modern political economy in the context of adverse selection, moral hazard, hold-up and other considerations, to derive optimal (in a second best sense) mechanisms to achieve public interest regulatory goals. This has become a vast literature; some of which is relevant to actual regulatory problems and practice, though much of it is not.

Let us start with the simplest characterization of the nature of the regulator's information disadvantages. A firm's costs may be high or low based on inherent attributes of its technical production opportunities, exogenous input cost variations over time and space, inherent differences in the costs of serving locations with different attributes (e.g. urban or rural), etc. While the regulator may not know the firm's true cost opportunities she will typically have some information about them. The regulator's imperfect information can be summarized by a probability distribution defined over a range of possible cost opportunities between some upper and lower bound within which the regulated firms actual cost opportunities lie. Second, the firm's actual costs will not only depend on its underlying cost opportunities but also on the behavioral decisions made by managers to exploit these cost opportunities. Managers may exert varying levels of effort to get more (or less) out of the cost opportunities that the firm has available to it. The greater the managerial effort the lower will be the firm's costs, other things equal. However, exerting more managerial effort imposes costs on managers and on society. Other things equal, managers will prefer to exert less effort than more to increase their own satisfac-

tion, but less effort will lead to higher costs and more "x-inefficiency." Unfortunately, the regulator cannot observe managerial effort directly and may be uncertain about its quality and impacts on the regulated firm's costs and quality of service.

The uncertainties the regulator faces about the firm's inherent cost opportunities gives the regulated firm a strategic advantage. It would like to convince the regulator that it is a "higher cost" firm that it actually is, in the belief that the regulator will then set higher prices for service as it satisfies the firm's long-run viability constraint (firm participation or budget-balance constraint), increasing the regulated firm's profits, creating dead-weight losses from (second-best) prices that are two high, and allowing the firm to capture social surplus from consumers. Thus, the social welfare maximizing regulator faces a potential *adverse selection* problem as it seeks to distinguish between firms with high cost opportunities and firms with low cost opportunities while adhering to the firm viability or participation constraint.

The uncertainties that the regulator faces about the quantity and impact of managerial effort creates another potential problem. Since the regulator typically has or can obtain good information about the regulated firm's actual costs (i.e. it's actual expenditures), at least in the aggregate, one approach to dealing with the adverse selection problem outlined above would simply be to set (or reset after a year) prices equal to the firm's realized costs ex post. This would solve the adverse selection problem since the regulator's information disadvantage would be resolved by auditing the firm's costs.[20] However, if managerial effort increases with the firm's profitability, this kind of "cost plus" regulation may lead management to exert too little effort to control costs, increasing the realized costs above their efficient levels. If the "rat doesn't smell the cheese and sometimes get a bit of it to eat" he may play golf rather than working hard to achieve efficiencies for the regulated firm. Thus, the regulator faces a potential *moral hazard* problem associated with variations in managerial effort in response to regulatory incentives [Laffont and Tirole (1986), Baron and Besanko (1987b)].

Faced with these information disadvantages, the social welfare maximizing regulator will seek a regulatory mechanism that takes the social costs of adverse selection and moral hazard into account, subject to the firm participation or budget-balance constraint that it faces, balancing the costs associated with adverse selection and the costs associated with moral hazard. The regulator may also take actions that reduce her information disadvantages by, for example, increasing the quality of the information that the regulator has about the firm's cost opportunities.

Following Laffont and Tirole [(1993, pp. 10–19)], to illuminate the issues at stake we can think of two polar case regulatory mechanisms that might be applied to a monopoly firm producing a single product. The first regulatory mechanism involves setting a fixed price ex ante that the regulated firm will be permitted to charge going forward. Alternatively, we can think of this as a pricing *formula* that starts with a particular price and

[20] Of course, the auditing of costs may not be perfect and in a multiproduct context the allocation of accounting costs between different products is likely to reflect some arbitrary joint cost allocation decisions.

then adjusts this price for *exogenous* changes in input price indices and other exogenous indices of cost drivers. This regulatory mechanism can be characterized as a *fixed price* regulatory contract or a *price cap* regulatory mechanism. There are two important attributes of this regulatory mechanism. Because prices are fixed (or vary based only on exogenous indices of cost drivers) and do not respond to changes in managerial effort, the firm and its managers are the residual claimants on production cost reductions and the costs of increases in managerial effort (and vice versa). That is, the firm and its managers have the highest powered incentives fully to exploit their cost opportunities by exerting the optimal amount of effort [Brennan (1989), Cabral and Riordan (1989), Isaac (1991), Sibley (1989), Kwoka (1993)]. Accordingly, this mechanism provides optimal incentives for inducing managerial effort and eliminates the costs associated with managerial moral hazard. However, because the regulator must adhere to a firm participation or viability constraint, when there is uncertainty about the regulated firm's cost opportunities the regulator will have to set a relatively high fixed price to ensure that if the firm is indeed inherently high cost, the prices under the fixed price contract or price cap will be high enough to cover the firm's costs. Accordingly, while the fixed price mechanism may deal well with the potential moral hazard problem by providing high powered incentives for cost reduction, it is potentially very poor at "rent extraction" for the benefit of consumers and society, potentially leaving a lot of rent to the firm due to the regulator's uncertainties about the firm's inherent costs and its need to adhere to the firm viability or participation constraint. Thus, while a fixed price contract solves the moral hazard problem it incurs the full costs of adverse selection.

At the other extreme, the regulator could implement a "cost of service" contract or regulatory mechanism where the firm is assured that it will be compensated for all of the costs of production that it incurs. Assume for now that this is a credible commitment—there is no ex post renegotiation—and that audits of costs are accurate. When the firm produces it will then reveal whether it is a high cost or a low cost firm to the regulator. Since the regulator compensates the firm for all of its costs, there is no "rent" left to the firm as excess profits. This solves the adverse selection problem. However, this kind of cost of service recovery mechanism does not provide any incentives for the management to exert effort. If the firm's profitability is not sensitive to managerial effort, the managers will exert the minimum effort that they can get away with. While there are no "excess profits" left on the table, consumers are now paying higher costs than they would have to pay if the firm were better managed. Indeed, it is this kind of managerial slack and associated x-inefficiencies that most policymakers have in mind when they discuss the "inefficiencies" associated with regulated firms. Thus, the adverse selection problem can be solved in this way, but the costs associated with moral hazard are fully realized.

Accordingly, these two polar case regulatory mechanisms each has benefits and costs. One is good at providing incentives for managerial efficiency and cost minimization, but it is bad at extracting the benefits of the lower costs associated with single firm production for consumers when costs are subadditive. The other is good at rent extraction but leads to costs from moral hazard. Perhaps not surprisingly, the optimal regulatory

mechanism (in a second best sense) will lie somewhere between these two extremes. In general, it will have the form of a *profit sharing* contract or a *sliding scale* regulatory mechanism where the price that the regulated firm can charge is *partially* responsive to changes in realized costs and *partially* fixed ex ante [Schmalensee (1989b), Lyon (1996)]. As we shall see, by offering a menu of regulatory contracts with different cost sharing provisions, the regulatory can do even better than if it offers only a single profit sharing contract [Laffont and Tirole (1993)]. The basic idea here is to make it profitable for a firm with low cost opportunities to choose a relatively high powered incentive scheme and a firm with high cost opportunities a relatively low-powered scheme. Some managerial inefficiencies are incurred if the firm turns out to have high cost opportunities, but these costs are balanced by reducing the rent left to the firm if it turns out to have low cost opportunities.

We can capture the nature of the range of options in the following fashion. Consider a general formulation of a regulatory process in which the firm's revenue requirements "R" are determined based on a fixed component "a" and a second component that is contingent on the firm's realized costs "C" and where "b" is the sharing parameter that defines the responsiveness of the firm's revenues to realized costs.

$$R = a + (1 - b)C$$

Under a fixed price contract or price cap regulation:

$$a = C^*$$

where C^* is the regulators assessment of the "efficient" costs of the highest cost type

$$b = 1$$

Under cost of Service regulation:

$$a = 0$$

$$b = 0$$

Under profit sharing contract or sliding scale regulation (Performance Based Regulation)

$$0 < b < 1$$

$$0 < a < C^*$$

These different mechanisms then have the properties summarized in Table 1.

The challenges then are to find the optimal performance based mechanism given the information structure faced by the regulator and for the regulator to find ways to reduce its information disadvantages vis a vis the regulated firm and to use the additional information effectively. As we shall see, it is optimal for the regulator to offer a menu of contracts with different combinations of a and b that meet certain conditions driven by the firm participation constraint and an incentive compatibility constraint that leads

Table 1
Incentives vs. Rent Extraction

Mechanism	Managerial Incentives	Rent Extraction
Fixed Price	100%	0%
Cost of Service	0	100%
Performance Based	$0 < x < 100\%$	$0 < y < 100\%$

firms with low cost opportunities to choose a high powered scheme (b is closer to 1 and a is closer to the efficient cost level for a firm with low cost opportunities) and firms with high cost opportunities to choose a lower powered incentive scheme (a and b are closer to zero). The lower powered scheme is offered to satisfy the firm participation constraint, sacrificing some costs associated with moral hazard, in order to reduce the rents that must be left to the high cost as it is induced to exert the optimal amount of managerial effort. (So far, this discussion has ignored quality issues. Clearly if a regulatory mechanism focuses only on reducing costs and ignores quality it will lead to firm to provide too little quality. This is a classic problem with price cap mechanisms and will be discussed further below.)

8.2. *Performance Based Regulation typology*

As I have already indicated, there is a very extensive theoretical literature on incentive regulation, or as it is commonly called by policymakers, performance based regulation or PBR. The papers that comprise this literature reflect a wide range of assumptions about the nature of the information possessed by the regulator and the firm about costs, cost reducing managerial effort, demand and product quality, the attributes of the regulatory instruments available to the regulator, the risk preferences of the firm, regulatory capture by interest groups, regulatory commitment, flexibility, and other dynamic considerations. These alternative sets of assumption can be applied in both a single or multiproduct context. One strand of the literature initially focused primarily on adverse selection problems motivated by the assumption that regulators could not observe a firm's costs and ignoring the role of managerial effort [Baron and Myerson (1982), Lewis and Sappington (1988a, 1988b)]. Another strand of the literature focused on both adverse selection and moral hazard problems motivated by the assumption that regulators could observe a firm's realized cost ex post, had information about the probability distribution of a firm's cost ex ante, and that managerial effort did affect costs but that this effort was not observable by the regulator [Laffont and Tirole (1986)]. Over time, these approaches have evolved to cover a similar range of assumptions about these basic information and behavioral conditions and lead to qualitatively similar conclusions. Armstrong and Sappington (2003a, forthcoming) provides a detailed and thoughtful review and synthesis of this entire literature and I refer readers interested in a very detailed treatment of the full range specifications of incentive regulation problems to their paper. Here I will simply lay out a "typology" of how these issues have been developed

in the literature and then provide some simple theoretical examples to illustrate what I consider to be the literature's primary conclusions of potential relevance for regulation in practice.

What are the regulator's objectives? Much of the literature assumes that the regulator seeks to maximize a social welfare function that reflects the goal of limiting the rents that are transferred from consumers and taxpayers to the firm's owners and managers subject to a firm participation or breakeven constraint. Armstrong and Sappington (2003a, 2003b) articulate this by specifying an objective function $W = S + \alpha R$ where W is expected social welfare, S equals expected consumers' (including consumers as taxpayers) surplus, R equals the expected rents earned by the owners and managers of the firm (over and above what is needed to compensate them for the total costs of production and the disutility of managerial effort to satisfy the participation constraint), and where $\alpha < 1$ implies that the regulator places more weight on consumers surplus than on rents earned by the firm. That is, the regulator seeks to extract rent from the firm for the benefit of consumers, subject as always to a firm participation or break-even constraint. In addition, W will be reduced if excessive rents are left to the firm since this will require higher (second-best) prices and greater allocative inefficiency.

Laffont and Tirole (1988a, 1988b, 1993, 2000) create a social benefit from reducing the rents left to the firm in a different way. In their basic model, consumer welfare and the welfare of the owners and managers of the firm are generally weighted equally. However, one of the instruments available to the regulator is the provision of transfer payments to the firm which affect the rents earned by the firm. These transfer payments come out of the government's budget and carry a social cost resulting from the inefficiencies of the tax system used to raise these revenues. Thus, for every dollar of transfer payments given to the firm to increase its rent, effectively $(1 + k)$ dollars of taxes must be raised, where k reflects the inefficiency of the tax system. Accordingly, by reducing the transfers to the firm over and above what is required to compensate it for its efficient production costs and the associated managerial disutility of effort, welfare can be increased. This set-up which allows for the use of costly government transfer payments also leads to a nice dichotomy between incentive arrangements that effectively establish the formula for determining the firm's revenues in a way that deals with adverse selection and moral hazard problems in the context of asymmetric information and price setting which establishes the second-best prices for the services sold by the firm given consumer demand attributes and the regulator's knowledge of them. That is, regulators first establish compensation arrangements (define how the firm's budget constraint or "revenue requirements" will be defined) to deal as effectively as possible with adverse selection and moral hazard problems given the information structure assumed. The regulator separately establishes a second best price structure to deal with allocational efficiency considerations which may not cover all of the firm's costs, with the difference coming from net government transfers. In addition, Laffont and Tirole (1993) introduce managerial effort as a variable that affects costs and service quality. Managers have a disutility of effort and must be compensated for it. Accordingly, the utility of management also appears in the social welfare function.

What does the regulator know about the firm ex ante and ex post? In what follows I will use the term "cost" to refer to the firm's marginal costs and ignore fixed costs (or normalize them to zero). This allows me to ignore in the discussion of incentive issues in this section, the second-best pricing (rather than incentive) options available to deal with budget-balance constraints created by increasing returns since the major issues associated with these pricing problems have been discussed above and are not affected in important ways by introducing asymmetric information about the firm's costs and managerial effort into the analysis. Carrying these issues forward here would simply complicate the presentation of the key incentive regulation results of interest. Accordingly, in what follows the full information benchmark is marginal cost pricing with zero rents left for the firm. Armstrong and Sappington (2003a, 2003b) distinguish between fixed costs and marginal costs, what the regulator knows about each and allow the regulator to make non-distortionary lump sum transfers to the firm. In this context, if the regulator can only make distortionary transfer payments, the full information benchmark with linear prices is Ramsey-Boiteux pricing and otherwise it is the optimal non-linear prices given the regulator's information about consumer demand.

The literature that focuses on adverse selection builds on the fundamental paper by Baron and Myerson (1982). The regulator does not know the firm's cost opportunities ex ante but has information about the probability distribution over the firm's possible cost opportunities.[21] Nor can the regulator observe or audit the firm's costs ex post. The firm does know its own cost opportunities ex ante and ex post. The firm's demand is known by both the regulator and the firm. There is no managerial effort in these early models. Accordingly, the analysis deals with a pure adverse selection problem with no potential inefficiencies or moral hazard associated with inadequate managerial effort. With moral hazard alone only high-powered incentive mechanisms are optimal. The regulator in the presence of adverse selection literature then proceeds to consider asymmetric information about the firm's demand function, where the firm knows its demand but either the regulator does not observe demand ex ante or ex post or learns about demand only ex post [Lewis and Sappington (1988a), Riordan (1984)]. Combining asymmetric information about costs and demand, introducing a multidimensional characterization of asymmetric information, is then a natural extension of the regulation to respond to adverse selection literature [Lewis and Sappington (1988b, 1989), Dana (1993), Armstrong and Rochet (1999)].

In light of common regulatory practice, a natural extension of these models is to assume that the regulated firm's actual realized costs are observable ex post, at least with uncertainty. Baron and Besanko (1984) considers cases where a firm's costs are "audited" ex post, but the actual realized costs resulting from the audit are observable by the regulator with a probability less than one. The regulator can use this information to reduce the costs of adverse selection. Laffont and Tirole (1986, 1993) consider cases

[21] In models that distinguish between fixed and variable costs, the regulator may know the fixed costs but not the variable costs. See Armstrong and Sappington (2003a, 2003b).

where the firm's realized costs are fully observable by the regulator. However, absent the simultaneous introduction of an uncertain scope for cost reductions through managerial effort, the regulatory problem then becomes trivial—just set prices equal to the firm's realized costs. Accordingly, Laffont and Tirole (1986a, 1993) introduce managers of the firm who can choose the amount of cost reducing effort that they expend. Managerial effort is not observable by the regulator ex ante or ex post, but realized production costs are fully known to the regulator as is the managerial "production function" that transforms managerial effort into cost reductions and the managers' utility over effort function. The regulated firm fully observes managerial effort, the cost reducing effects of managerial effort, and demand. It also knows what managerial utility would be at different levels of effort. Armstrong and Sappington (2003a, 2003b) advance this analysis by considering cases where the regulated firm is uncertain about the operating costs that will be realized but knows that it can reduce costs by increasing managerial effort, though in a way that creates a moral hazard problem but no adverse selection problem. In the face of uncertainty over its costs, they consider cases where the firm may be either risk-neutral or risk averse.

The literature also examines situations in which the regulator is *captured* by an interest group and no longer seeks to maximize social welfare W. For example, the regulator may be bribed not to use or reveal information that would reduce the rents available to the firm [Laffont and Tirole (1993, Chapter 11)] and the regulator may effectively collude with the firm if she can be compensated in some way (monetary, future employment, jobs for friends and relatives) for doing so. The possibility of regulatory capture may affect the choice of the power of the incentive schemes used by the regulator. High powered incentive schemes are more susceptible to regulatory capture than are lower powered schemes [Laffont and Tirole (1993, pp. 57–58)]. To counteract the possibility of regulatory collusion, the analysis can also be expanded to include another level in which the government imposes an incentive scheme on the regulator to provide incentives to reveal and use all relevant information possessed by the regulator and more generally not to collude with interest groups.

What instruments are available to the regulator and how do the regulator and the regulated firm interact over time? Much of the incentive regulation literature is static. The regulator (or the government through the regulator) can offer a menu of prices (or fixed price contracts) with or without a fixed fee or transfer payment. The menu may contain prices that are contingent on realized costs (which can be thought of as penalties or rewards for performance) in those models where regulators observe costs ex post. Some of these instruments may be costly to utilize (e.g. transfer payments and auditing efforts). The more instruments the regulator has at its disposal and the lower the costs of using them, the closer the regulator will be able to get to the full information benchmark.

Of more interest are issues that arise as we consider the dynamic interactions between the regulated firm and the regulator and the availability and utilization of mechanisms that the regulator potentially has available to reduce its information disadvantage. It is inevitable that the regulator will learn more about the regulated firm as they interact over

time. So, for example, if the regulator can observe a firm's realized costs ex post, should the regulator use that information to reset the prices that the regulated firm receives [commonly known as a "ratchet"—Weitzman (1980)]? Or is it better for the regulator to commit to a particular contract ex ante, which may be contingent on realized costs, but the regulator is not permitted to use the information gained from observing realized costs to change the terms and conditions of the regulatory contract offered to the firm? Is it credible for the regulator to commit *not* to renegotiate the contract, especially in light of U.S. regulatory legal doctrines that have been interpreted as foreclosing the ability of a regulatory commission to bind future commissions?

Clearly, if the regulated firm knows that information about its realized costs can be used to renegotiate the terms of its contract, this will affect its behavior ex ante. It may have incentives to engage in less cost reduction in period 1 or try to fool the regulator into thinking it is a high cost firm so that it can continue to earn rents in period 2. Of if the regulated firm has a choice between technologies that involve sunk cost commitments, will the possibility of ex post opportunism or regulatory expropriation, perhaps driven by the capture of the regulator by other interest groups, affect its willingness to invest in the lowest cost technologies when they involve more significant sunk cost commitments (leading to the opposite of the A-J effect).

These dynamic issues have been examined more intensively over time and represent a merging of the literature on regulation with the literature on contracts and dynamic incentive mechanisms more generally [Laffont and Tirole (1988b, 1990a, 1993), Baron and Besanko (1987a), Armstrong and Vickers (1991, 2000), Armstrong, Cowan, and Vickers (1994)]. The impacts of regulatory lag of different durations [Baumol and Klevorick (1970), Klevorick (1973), Joskow (1974)] and other price adjustment procedures have been analyzed extensively as well [Vogelsang and Finsinger (1979), Sappington and Sibley (1988, 1990)].

8.3. Some examples of incentive regulation mechanism design

This section is based on Laffont and Tirole (1993, Chapter 2). We will examine the case of a regulated monopoly firm producing a single private good and which is restricted to charging linear prices. A specific firm's cost opportunities depend on the best technology and input prices that it has access to and which will characterize it's "type" denoted by β. The firm knows its type but the regulator is uncertain about the firm's type. We will begin with a two-type case where the firm can be either a low cost type denoted by β_L with probability v or a high cost type β_{H} with probability $1 - v$. The firm's management can exert effort e but managerial utility declines as effort increases. The firm's cost function is then given by:

$$C(q) = F + (\beta - e)q$$

Assume that F is known by the regulator and we normalize it to zero for simplicity. The regulator cannot observe β or e, but can observe the firm's actual production costs ex

post. Then the firm's marginal cost is given by

$$c = (\beta - e)$$

and the disutility of managers with respect to effort e is defined as

$$U = t - \psi(e)$$

where $\psi'(e) > 0$. This function is known to the firm and to the regulator, but the regulator cannot observe e or U directly.

Define:

$S(q) =$ gross consumers' surplus

$q = D(p) =$ market demand curve for the product

$P = P(q) =$ inverse market demand curve for the product

$R(q) = qP(q) =$ market revenue generated by the firm

Laffont and Tirole (1993) allow the government to make financial transfers to the firm with a social cost of λ per dollar transferred, so that to transfer one dollar to the firm costs the government (and society) $(1 + \lambda)$ dollars. To keep the accounting straight we adopt Laffont and Tirole's accounting convention. All revenues from sales of the product go to the government and then the government reimburses the firm for its actual production costs plus an additional transfer payment that is greater than or equal to zero. Thus, the firm's costs are covered (breakeven constraint is satisfied) and the (net) transfer payment t must be large enough at least to compensate the managers for the disutility of effort to satisfy the participation constraint. Then social welfare W is given by:

$$\begin{aligned} W &= V(q) - (1+\lambda)(C + t) + U \\ &= V(q) - (1+\lambda)(C + \psi(e)) - \lambda U \end{aligned} \tag{34}$$

where:

$$\begin{aligned} V(q) &= \big[S(q) - R(q)\big] + (1+\lambda)R(q) \\ &= S(q) + \lambda q P(q) \end{aligned}$$

The full information benchmark is then derived as follows:

$$\max_{(e,q)} W = S(q) + \lambda q P(q) - (1+\lambda)\big[(\beta - e)q + \psi(e)\big] - \lambda U \quad \text{s.t. } U \geq 0 \tag{35}$$

The first order conditions are:

$$U = 0 \text{ [no rent left to the firm/managers, but participation constraint is satisfied]} \tag{36}$$

$$\psi'(e) = q \text{ [marginal disutility of effort equals marginal cost savings from additional effort]} \tag{37}$$

$$P(q) + \lambda P(q) + \lambda q P'(q) = (1 + \lambda)(\beta - e) = (1 + \lambda)c$$

or

$$(p(q) - c)/p(q) = \left[\lambda/(1 + \lambda)\right](1/\eta) \text{ [Ramsey-Boiteux pricing]} \tag{38}$$

where η is the elasticity of demand for the product supplied by the regulated firm.

Condition (38) requires some explanation. It looks like the Ramsey-Boiteux pricing formula that we discussed earlier and, in a sense, it is. However, here λ is not the shadow price of the firm's budget constraint but rather the marginal cost of raising government revenues through the tax system and then distributing government revenues to the firm to cover its costs and a transfer payment to compensate managers for their disutility of effort. The optimal prices here serve a pure social allocation function that take into account the cost of using public funds to compensate the firm for its costs and managers for their disutility of effort. These "Ramsey-Boiteux prices" are equivalent to adding the optimal commodity taxes to the marginal cost of supplying these services. This is the essence of Laffont and Tirole's separation of or dichotomy between "incentives to deal with moral hazard and adverse selection" and "prices" to deal with consumption allocational considerations.

To summarized, with full information, the regulator would compensate the firm for its costs and the manager's disutility of effort leaving no rents to the firm (36). It would also require the managers of the firm to exert the optimal effort e^*, which in turn yields the optimal level of total and marginal costs (b). Let $q^*(c)$ denote the solution to (37) and (38) and call it the Ramsey-Boiteux output. Then $P(q^*(c))$ is the Ramsey-Boiteux price.

Now we consider the characteristics of (second-best) optimal regulatory mechanism when there is asymmetric information. Everything is common knowledge except the regulator cannot observe the firm's type β or the quantity of managerial effort e expended. In the most simple case, the regulator does know that the firm is either a high cost type with β_L with probability v or a high cost type β_{H} with probability $(1 - v)$. The attributes of the optimal regulatory mechanism are then derived by maximizing expected social welfare given the probability of each type subject to a firm viability constraint ($U \geq 0$) for each type and an incentive compatibility constraint that ensures that each type chooses the regulatory contract that is optimal given asymmetric information. Laffont and Tirole (1993) show that the binding incentive compatibility constraint is given by the low-cost type's rent which in turn is determined by the high cost type's marginal cost. Basically, the contract designed for the high cost type leaves no rent to the high-cost firm and its managers while the contract designed for the low cost type must leave enough rent to the low cost type so that it does not choose the contract designed for the high cost type. This rent is the difference in their realized marginal costs at the effort levels they choose given the contract they take up.

The expected welfare seen by the regulator is

$$\begin{aligned}\underset{(U_L,U_{\mathrm{H}},q_{\mathrm{H}},q_L,e_{\mathrm{H}},e_L)}{\operatorname{Max}} W &= v\big[V(q_L) - (1+\lambda)\big[(\beta_L - e_L)q + \psi(e_L)\big] - \lambda\Phi(e_{\mathrm{H}})\big] \\ &\quad + (1-v)\big[V(q_{\mathrm{H}}) - (1+\lambda)\big[(\beta_{\mathrm{H}} - e_{\mathrm{H}})q_{\mathrm{H}} + \psi(e_{\mathrm{H}})\big]\big]\end{aligned} \tag{39}$$

(subject to firm participation constraints and incentive compatibility constraints) where $\Phi(\cdot)$ is an increasing function of $e = \psi(e) - \psi(e - (\beta_{\mathrm{H}} - \beta_L))$. Maximizing expected welfare subject to the firm participation and incentive compatibility constraints yields the first order conditions are:

$$q_L = q^*(\beta_L - e_L) \tag{40}$$

$$\psi'(e_L) = q_L \tag{41}$$

$$q_{\mathrm{H}} = q^*(\beta_{\mathrm{H}} - e_{\mathrm{H}}) \tag{42}$$

$$\psi'(e_{\mathrm{H}}) = q_{\mathrm{H}} - \big[\lambda/(1+\lambda)\big]\big[v/(1-v)\big]\Phi'(e_{\mathrm{H}}) \tag{43}$$

First order conditions (41) and (42) are simply the Ramsey-Boiteux quantities given the realization of marginal cost and the associated Ramsey-Boiteux prices are optimal for each type. That is, they are the same as under full information. First order condition (41) shows that the optimal contract for the low-cost type will induce the low cost type to exert the optimal amount of effort as it would under full information. First order condition (43) shows that the effort exerted by the high cost type will be less than optimal. The firm participation constraint is also binding for the high cost type ($U = 0$) but not for the low-cost type ($U > 0$). Thus, while the low cost type chooses the optimal amount of effort, it gains an information rent $U > 0 = \Phi(\beta_{\mathrm{H}} - \beta_L)$. The reason that the effort of the high cost type is optimally distorted from the full information optimal level is to reduce the rent that must be left to the low cost type to satisfy the incentive compatibility constraint which is binding for the high cost type. Reducing e_{H} by a small amount has two effects. It reduces the disutility of effort and increases the cost of production. The net effect on the firm's unit cost, including managerial disutility of effort, is $1 - \psi(e_{\mathrm{H}})$. But this also reduces the rent that must be left to the low cost firm by $\Phi'(e_{\mathrm{H}})$ to satisfy the incentive compatibility constraint. So the expected increase in the net unit cost to the high cost firm are $(1-v)(1+\lambda)(1-\psi(e_{\mathrm{H}}))$ and the reduction in the unit cost of rent transfers to the low cost firm is $v\lambda\Phi'(e_L)$. The amount of the distortion in e_L is then chosen to equate these costs on the margin.

The optimal regulatory mechanism involves offering the regulated firm a choice between two regulatory contract options. One is a fixed price option that leaves some rent if the firm is a low-cost type but negative rent if it is a high cost type. The second is a cost-contingent contract that distorts the firm's effort if it is a high cost type but leaves it no rent. The high powered scheme is the most attractive to the low-cost type and the low-powered scheme is the most attractive to the high cost type. The expected cost of the distortion of effort if the firm is a high cost type is balanced against the expected cost of leaving additional rent top the firm if it is a low cost type—*the fundamental tradeoff between incentives and rent extraction*.

The two-type example can be generalized to a continuum of types [Laffont and Tirole (1993, pp. 137ff)]. Here we assume that β has a continuous distribution from some lower bound β_L to some upper bound β_H with a cumulative distribution $F(\beta)$ and a strictly positive density $f(\beta)$ where F is assumed to satisfy a monotone hazard rate condition and $F(\beta)/f(\beta)$ is non-decreasing in β. The regulator maximizes expected social welfare subject to the firm participation and incentive compatibility constraints as before and incentive compatibility requires a mechanism that leaves more rent to the firm the lower is its type β, with the highest cost type getting no rent, the lowest cost type getting the most rent and intermediate type's rent defined by the difference in their marginal costs. Similarly, the effort of the lowest cost type is optimal and the effort of the highest cost type is distorted the most, with intermediate types having smaller levels of distortion (and more rents) as β declines toward β_L. In the case of a continuous distribution of types, the optimality conditions are directly analogous to those for the two-type case.

$$q(\beta) = q^*\big(\beta - e(\beta)\big) \quad \text{[Ramsey Pricing]} \tag{44}$$

$$\Psi'(e(\beta)) = q(\beta) - \big[\lambda/(1+\lambda)\big]\big[F(\beta)/f(\beta)\big]\Psi''(e(\beta)) \tag{45}$$

Where (44) shows that Ramsey pricing is optimal given realized costs and (45) shows that effort is distorted as β increases to constrain the rents that are left to lower cost firms.

Laffont and Tirole (1993) show that these optimality conditions can be implemented by offering the firm a menu of linear contracts, which in their model are transfer or incentive payments in excess of realized costs (which are also reimbursed), of the form:

$$t(\beta, c) = a(\beta) - b(\beta)c$$

where a is a fixed payment, b is a cost contingent payment, and a and b are decreasing in β.

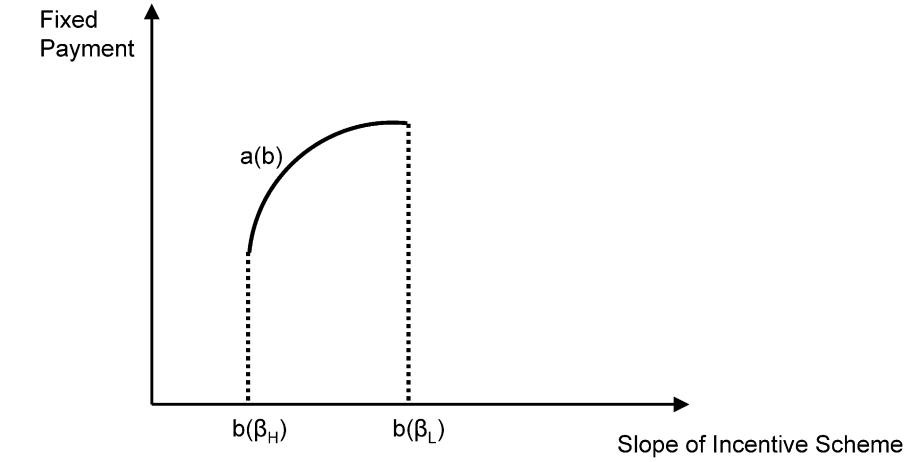
We can rewrite the transfer payment equation in terms of the gross transfer to the firm including the unit cost reimbursement:

$$R_f = a(\beta) - b(\beta)c + c = a(\beta) + \big(1 - b(\beta)\big)c \tag{46}$$

where $\mathrm{d}a/\mathrm{d}b > 0$ (for a given β a unit increase in the slope of the incentive payment must be compensated by an increase in the fixed payment to cover the increase in production costs) and $\mathrm{d}^2a/\mathrm{d}b^2 < 0$ (the fixed payment is a concave function of the slope of the incentive scheme).

Figure 11 displays this relationship. The lowest cost type chooses a fixed price contract with a transfer net of costs equal to U_L and the firm is the residual claimant on cost reducing effort ($b = 1$). As β increases, the transfer is less sensitive to the firm's realized costs (b declines) and the rent is lower (a declines).

Note that if one were to try empirically to relate the firms' realized costs to the power of the incentive scheme they had selected, a correlation between the power of the contract and the firm's realized costs would not tell us anything directly about the



Source: Laffont and Tirole (1993), Figure 1.5

Figure 11. Menu of incentive contracts.

incentive effects of higher-powered schemes in terms of inducing optimal effort and mitigating moral hazard problems. This is the case because the firms with the lower inherent costs will rationally choose the higher powered contracts. Assume that we had data for regulated firms serving different geographic regions (e.g. different states) which had different inherent cost opportunities (a range of possible values for β). If the regulators in each state offered the optimal menu of incentive contracts, the low β firms would choose high powered contracts and the high β firms would choose lower powered contracts. Accordingly, the effects of the mechanisms on mitigating the rents that would accrue to a low cost firm's information advantage from the effects on inducing optimal effort are not easily distinguished. I know of only one empirical paper that has endeavored to tackle this challenge directly [Gagnepain and Ivaldi (2002)] and it is discussed further below.

8.3.1. The value of information

This framework also provides us with insights into the value to the regulator of reducing her information disadvantage. Consider the two-type case. Let's say that the regulator is able to obtain information that increases her assessment of the probability that the firm is a low cost type from v to v_{H}. If the regulator's assessment of v increases there are two effects. The first effect is that the rent left to the low cost type falls. By increasing v, more weight in the social welfare function is placed on the realization of the firm being a low cost type and this increases the expected cost of rent transfers other things equal. Similarly, the optimal distortion induced in the high-cost type increases since less weight is placed on this realization in the expected welfare function. These

intuitive results carry through for the continuous type case. Overall, as the regulator's information becomes more favorable as defined in Laffont and Tirole (1993, pp. 76–81), the higher is welfare even though for a given realization of β we will observe firm's choosing (being offered) a lower powered incentive scheme. This latter result, does not appear to be generalizable to models where there are no government transfers and where revenues the firm earns from sales must be relied upon entirely to achieve both incentive goals (adverse selection and moral hazard) and allocational goals [Armstrong, Cowan, and Vickers (1994, pp. 39–43), Schmalensee (1989b)]. Without government transfers as an instrument, if the regulator's uncertainty about firm types declines she will choose a higher powered scheme, because the budget balance constraint effectively becomes less binding, allowing the regulator to tolerate more variation in a firm's realized net revenues.

One way in which regulators can effectively reduce their information advantage is by using competitive benchmarks or "yardstick regulation" in the price setting process. Schleifer (1985) shows that if there are $n > 1$ non-competing but otherwise identical firms (e.g. gas distribution companies in firms in different states), an efficient regulatory mechanism involves setting the price for each firm based on the costs of the other firms. Each individual firm has no control over the price it will be allowed to charge (unless the firms can collude) since it is based on the realized costs of $(n - 1)$ other firms. So, effectively each firm has a fixed price contract and the regulator can be assured that the budget balance constraint will be satisfied since if the firms are identical prices will never fall below their "efficient" realized costs. This mechanism effectively induces each firm to compete against the others. The equilibrium is a price that just covers all of the firm's efficient costs as if they competed directly with one another.

Of course, it is unlikely to be able to find a large set of truly identical firms. However, hedonic regression, frontier cost function estimation and related statistical techniques can be used to normalize cost variations for exogenous differences in firm attributes to develop normalized benchmark costs [Jamasb and Pollitt (2001, 2003), Estache, Rossi, and Ruzzier (2004)]. These benchmark costs can then be used by the regulator in a yardstick framework or in other ways to reduce its information advantage, allowing it to use high powered incentive mechanisms without incurring the cost of excessive rents that would accrue if the regulator had a greater cost disadvantage.

Laffont and Tirole (1993, pp. 84–86) offer a simple model that characterizes the issues as stake here. Let's say that the regulator is responsible for two non-competing firms ($i = 1, 2$) that each produce one unit of output supplied in separate geographic areas. Their costs are given by:

$$C^i = \beta^{\mathrm{a}} + \beta^i - e^i$$

Where β^{a} is an aggregate shock to both firms and β^i is an idiosyncratic shock that is independent of β^j and e^i is firm i's effort. As before the firm's rent is given by

$$U^i - \psi(e^i)$$

and the regulator can observe only realized costs. Each firm learns the realizations of its own shocks before choosing from the regulator contracts offered to it. Laffont and Tirole develop several cases:

Case 1: In the case of purely idiosyncratic shocks ($\beta^{\mathrm{a}} = 0$), the firms are unrelated and we are back to the standard case where they must be regulated separately.

Case 2: In the case of purely aggregate shocks ($\beta^{i} = 0$) the regulator can achieve the first best outcome by offering the firms only a fixed price contract based on their relative performance or "yardstick regulation." The transfer or incentive payment is then given by $t^{i} = \Psi(e^{*}) - (C^{i} - C^{j})$. Firm i maximizes $\{\Psi(e^{*}) - [(\beta^{\mathrm{a}} - e^{i}) - (\beta^{\mathrm{a}} - e^{j})]\} - \Psi(e^{i})$ and chooses $e^{i} = e^{*}$. Since the other firm is identical it also chooses e^{*}. Neither firm earns any rents and they both exert the optimal amount of effort (they are identical). By filtering aggregate uncertainty out of each firm's realized costs we can get to the first best.

Case 3: In the case of general shocks that cannot be separated into aggregate and idiosyncratic components, a mechanism can be designed that is based in part on relative performance that has superior welfare properties to the Laffont-Tirole menu of contracts

8.3.2. Ratchet effects or regulatory lag

So far, this analysis assumes the regulator establishes a regulatory contract once and for all. This assumption is important for the results because it is assumed that the regulator can observe the firm's realized costs ex post. If the regulator then used this information to reset the firm's prices, the firm would have a less powerful incentive to engage in cost reducing effort—a "ratchet" (Weitzman, 1980). More generally, as we discussed earlier, the behavior of a firm will depend on the information that its behavior reveals to the regulator ex post and how the regulator uses that information in subsequent regulatory reviews. The effects of this kind of interaction between the regulated firm and the regulator can be captured in the Laffont and Tirole model in a straightforward manner [Laffont and Tirole (1993, Chapter 9)].

Consider the two-type case again and ignore discounting. The low cost type will choose the high powered incentive contract and will earn a rent of $\Phi(e_{\mathrm{H}})$ until the regulator resets its prices to equal its realized costs at which time its rents will fall to zero. This is not incentive compatible. The low cost type would do better by exerting less effort in the first period, reducing its disutility of effort, leading its realized production costs to increase, effectively mimicking the observable production costs expected by the regulator for the high cost type (effectively leading the regulator to believe incorrectly that the low cost type is a high cost firm). The low cost firm still earns rents in period 1, but through a lower disutility of effort. Post-ratchet, the firm faces a fixed price set equal to its realized production costs in period 2 and can now exert optimal effort and earn rents again post-ratchet by reducing its production costs.

To restore incentive compatibility with a ratchet, the low-cost type would have to be given a larger rent in period 1, at least as large as the rent it can get in period 2 after mimicking the production costs of the high cost type in period 1. However, if the first

period rents are high enough, the high cost firm may find it attractive to choose the high powered incentive scheme in period 1 and then go out of business in period 2. Laffont and Tirole call this the "take the money and run" strategy.

These simple examples are obviously rather contrived. However, we can find examples of them in the real world. The regulatory mechanism utilized extensively in the U.K. since its utility sectors were privatized is effectively a fixed price contract (actually a price cap that adjusted for general movements in input prices and an assumed target rate of productivity growth—a so-called RPI-X mechanism as discussed further below) with a ratchet every five (or so) years when the level of the price cap is reset to reflect the current realized (or forecast) cost of service [Beesley and Littlechild (1989), Brennan (1989), Isaac (1991), Sibley (1989), Armstrong, Cowan, and Vickers (1994), Joskow (2005a, 2005b)]. It has been observed that regulated firms made their greatest cost reduction efforts during the early years of the cap and then exerted less effort at reducing costs as the review approached [OFGEM (2004a, 2004b)]. More generally, the examples make the important point that the dynamic attributes of the regulatory process and how regulators use information about costs revealed by the regulated firm's behavior over time have significant effects on the incentives the regulated firm faces and on its behavior [Gilbert and Newbery (1994)].

8.3.3. No government transfers

How do the basic results developed with the Laffont-Tirole framework change if no government transfers are permitted? Clearly, regulated prices alone must now serve to deal with adverse selection, moral hazard, and allocational issues. The dichotomy between prices and incentives no longer holds. However, the same basic attributes of incentive contracting continue to apply. Focusing on linear pricing, it is optimal for the regulator to offer a menu of cost contingent price options—cost sharing or sliding scale contracts—where the attributes of the menu are chosen to balance the firms budget (production cost plus incentive payment) in a way that trades off rent extraction, effort incentives, and allocation distortions subject to participation and incentive compatibility constraints [Laffont and Tirole (1993, pp. 151–153)]. The lowest price in the menu is a fixed price designed to be chosen by the low-cost opportunity firm. The price gives that firm high powered incentives to exert cost-reducing effort, but it also leaves the most rent to the firm and involves the greatest departure of price from marginal cost. As we move to higher cost types the price increases as does the sensitivity of the price level to changes in costs. Incentives for cost reducing effort decline as β increases, rents left to the firm fall, and prices are closer to the firm's realized marginal cost, though this cost is too high due to suboptimal effort.

8.4. Price regulation when cost is not observable

As noted earlier, the earliest modern theoretical work on incentive regulation [Baron and Myerson (1982)] assumed that the regulator could not observe costs at all, could

observe demand, and that there were no moral hazard problems.[22] The regulator cares about rent extraction and must adhere to a firm participation or viability constraint. In this context, the regulatory problem is an adverse selection problem and cost contingent contracts are not available instruments since costs are assumed not to be observable. I do not find this to be a particularly realistic characterization of regulation in many developed countries, but especially in developing countries, regulators often have difficulty getting credible cost information from the firms they regulate. Moreover, accurate and meaningful cost measurement may be very difficult for multiproduct firms that have joint costs. Accordingly, I will conclude this section with a brief discussion of this literature.

Let us begin with Baron and Myerson (1982), using the development in Laffont and Tirole (1993, pp. 155–158)]. Consider a firm that has cost

$$C = \beta q$$

where the regulator observes q, has a probability distribution over β, there is no moral hazard (e), and the firm receives revenues from sales at the regulated price P and a transfer payment t from the regulator. The firm's utility is now

$$U = t + P(q)q - \beta q$$

where $P(q)$ is the inverse demand function. Since cost is not observable, the regulator must rely on fixed price contracts (and accordingly if we added moral hazard the firm would have optimal cost-reducing incentives).

If the regulator had full information the optimal linear price would be the Ramsey-Boiteux price and the associated Ramsey-Boiteux output:

$$L = (p - \beta)/p = (\lambda/1 + \lambda)(1/\eta) \tag{47}$$

where λ is either the shadow cost of public funds with government transfers or the shadow price of the budget constraint when the firm must balance its budget from sales revenues and there are fixed costs to cover.

With asymmetric information of the kind assumed here, the regulator will offer a menu of fixed price contracts that are distorted away from the Ramsey-Boiteux prices given β. The distortion for a given β reflects the tradeoff between the allocational distortion from increasing prices further above marginal cost and the cost of leaving more rent to the firm subject to the firm viability and incentive compatibility constraints.

$$L = \big(p(\beta) - \beta\big)/p(\beta) = (\lambda/1 + \lambda)(1/\eta) + (\lambda/1 + \lambda)\big[(F(\beta)/(f(\beta)p(\beta)\big] \tag{48}$$

[22] Loeb and Magat (1979) propose a mechanism where the regulator can observe the firm's demand function and can observe price and quantity ex post. The regulator does not care about the distribution of the surplus. They propose a mechanism that offers the regulated firm a subsidy equal to the total consumer surplus ex post. The firm then has an incentive to set price equal to marginal cost to maximize its profits. If the regulator cares about the distribution of income (rent extraction) it could auction of this regulatory contract in a competition for the market auction. The mechanism then reduces to Demsetz's franchise bidding scheme. The latter raises numerous issues that are discussed above.

where $F(\beta)$ and $f(\beta)$ are as defined before and $\mathrm{d}[F(\beta)/(f\beta)]\mathrm{d}\beta \geq 0$. Prices clearly exceed the Ramsey-Boiteux level at all levels of β (compare (47) and (48)). Absent the ability to use a cost-contingent reimbursement mechanism, prices must be distorted away from their Ramsey levels to deal with rent extraction/adverse selection costs.

This analysis has been extended by Baron and Besanko (1984) to allow for random audits of the firm's costs by the regulator, again in the absence of moral hazard. The firm announces its costs and prices ex ante and there is some probability that the firm will be audited ex post. The result of the audit is a noisy measure of the firm's actual costs. After the audit the regulator can penalize the firm if it gets a signal from the audit that the firm's actual costs are greater than its announced costs. Absent moral hazard the optimal policy is to penalize the firm when the audit yields a measured cost that is low, signaling the regulator that the firm's costs are likely to be lower than the cost and associated price that the firm announced. (With moral hazard things are more complicated because one does not want to penalize the firm for cost-reducing effort.) The threat that the firm's announced costs will be audited reduces the price the firm charges, the rents it retains, and the allocational distortion from prices greater than costs.[23]

8.5. *Pricing mechanisms based on historical cost observations*

Vogelsang and Finsinger (1979) have developed a mechanism that relies on observations of a regulated firm's prices, output and profits to adjust the firm's prices over time. Their mechanism, as characterized by Laffont and Tirole [Laffont and Tirole (1993, pp. 162–163)], gives the firm a reward or bonus at each point in time defined by

$$B_t = a + (\Pi_t - \Pi_{t-1}) + (p_{t-1} - p_t)q_{t-1} \tag{49}$$

where $\Pi_t = (p_t q_t - C(q_t))$. Basically, the mechanism rewards price reductions up to a point. Think of p_t as starting at the monopoly price. If the firm leaves its price at this level it gets a bonus payment of a. If it reduces its price in the second period, q will increase, profits will fall from t to $t-1$, but total revenue will increase since $\mathrm{MR} > 0$ at the monopoly price. The increase in revenue from the second term will exceed the reduction in profits from the first term, increasing net profit under the bonus formula when the price falls from the pure monopoly level and the bonus will be higher than if the firm left its price at the monopoly level. The regulator is bribing the firm to lower its prices in order to reduce the allocative distortions from prices that are too high by rewarding it with some of the increases in infra-marginal consumers surplus resulting from lower prices. Finsinger and Volgelsang show that the firm has the incentive to continue to reduce its price until it reaches the Ramsey-Boiteux price. However, if cost reducing effort is introduced, the cost-contingent nature of this mechanism leads to too little cost reducing effort [Sappington (1980), Laffont and Tirole (1993, pp. 142–145)].

[23] Lewis and Sappington (1988a) extend this line of attached to assume that the firm has private information about demand rather than costs and extend the analysis [Lewis and Sappington (1988b)] to assume private information about both demand and costs.

9. Measuring the effects of price and entry regulation

Price and entry regulation may affect several interrelated performance indicia. These indicia include the level of prices, the structure of price charged to different groups of customers or for different products, prices for inputs paid by the regulated firm, the firm's realized costs of production, firm profits, research and development activity, the adoption of product and process innovations, and the distribution of economic rents between shareholders, consumers and input suppliers. To measure the effects of regulation one must first decide upon the performance norms against which regulatory outcomes are to be measured. Candidate benchmarks include characterizations or simulations of fully efficient outcomes, hypothetical unregulated/competitive outcomes, and outcomes resulting from the application of alternative regulatory mechanisms. The identification of benchmarks and, especially the use of alternative benchmarks for normative evaluation of the effects of regulatory mechanisms and processes, should be sensitive to the fact that fully efficient outcomes or perfectly competitive outcomes are unlikely to be achievable in reality. Accordingly, what Williamson (1996, pp. 237–238) refers to as a *remediableness* criterion should be applied in normative evaluations. That is, what is the best that can be done in an imperfect world?

Once the relevant benchmarks have been identified there are several different empirical approaches to the measurement of the effects of price and entry regulation on various performance indicia.

1. *Cross-sectional/Panel-data analysis:* These studies examine the performance indicia for firms serving different geographic areas and subject to different intensities or types of regulation, typically measured over a period of more than one year. For example studies may compare prices, costs, profits, etc. for similar firms serving customers in different states under different regulatory regimes for a period of one or more years. The classic study here is that of Stigler and Friedland (1962) where they examined differences in electricity prices between states with commission regulation and states without state commission regulation of electricity prices. Or the cross-sectional variation may be between states that use different types of regulatory mechanisms (traditional cost of service with or without PBR enhancements) or apply similar mechanisms more or less intensively [Mathios and Rogers (1989)]. The assumption in many of these studies is that the choice of whether to regulate or not is exogenous, so that cross-sectional data provide observations of "natural experiments" in the impacts of the effects of alternative regulatory mechanisms. However, several recent panel data studies recognize that the choice of regulatory instrument may be endogenous [e.g. Ai and Sappington (2002), Sappington and Ai (2005)].

   Natural or near-natural experiments that produce cross-section and time series variations in the nature or intensity of regulatory mechanisms can and have provided very useful opportunities to measure the effects of regulation, the effects of variations in the structure of regulatory mechanisms and the impacts of dereg-

ulation initiatives. However ensuring that one really has a meaningful natural experiment is always a challenge [Joskow (2005b)].

In principle, cross-country comparisons can be used in an equivalent fashion, though differences in accounting conventions, data availability, and basic underlying economic and institutional attributes make cross-country studies quite difficult. Nevertheless, there has been increasing use of cross-country data both to evaluate the effects of regulation and to provide data to develop performance benchmarks that can be used by regulators [Carrington, Coelli, and Groom (2002), Jamasb and Pollitt (2003)].

2. *Time series or "before and after" analysis:* These studies measure the effects of regulation by comparing various performance variables "before" and "after" a significant change in the regulatory environment. Much has been learned about the effects of price and entry regulation by comparing firm and industry behavior and performance under regulation with the changes observed after price and entry regulations are removed. Or it could be a shift from cost of service regulation to price cap regulation. Here the challenge is to control for other factors (e.g. input prices) that may change over time as well and the inconvenient fact that regulation and deregulation initiatives are often phased in over a period of time and not single well defined events.
3. *Structural models and policy simulations:* These studies specify and estimate the parameters of firm and/or industry demand, firm costs, and competitive interactions (if any) to compare actual observations on prices, costs, profits, etc. with simulated prices under alternative regulated and unregulated regimes. This is most straightforward in the case of legal monopolies. With the demand and cost functions in hand, the optimal prices can be derived and compared to the actual prices. Or industries where there are multiple firms competing based on regulated prices, the actual prices can be compared to either optimal prices or "competitive" prices, once the nature of competitive interactions have been specified. Related work measures the attributes of production functions and tests for cost minimization. Still other work uses models of consumer demand to measure the value of new products and, in turn, the social costs of regulatory delays in the introduction. Related work on process innovations can also be incorporated in a production function framework for similar types of analysis.

9.1. Incentive regulation in practice

There is an extensive literature that examines empirically the effects of price and entry regulation in sectors in which there are (or were) legal monopolies to serve specific geographic areas (e.g. electricity, gas distribution, water, telephone) as well as in sectors in which prices and entry were regulated but two or more firms were given the legal authority to compete in the market (e.g. airlines, trucking, railroads, automobile insurance, natural gas pipelines). Reviews of the pre-1990 literature on the measurement of the effects of regulation in both single and multi-firm settings can be found in Joskow

and Noll (1981), Berg and Tschirhart (1988), Joskow and Rose (1989), Winston (1993), (Joskow, 2005b). I will focus here on the more recent nascent literature that examines the effects of incentive or performance based regulation of legal monopolies.

Although the theoretical literature on incentive regulation is fairly recent, we can trace the earliest applications of incentive regulation concepts back to the early regulation of the gas distribution sector[24] in England in the mid-19th century [Hammond, Johnes, and Robinson (2002)]. A sliding scale mechanism in which the dividends available to shareholders were linked to increases and decreases in gas prices from some base level was first introduce in England in 1855 [Hammond, Johnes, and Robinson (2002, p. 255)]. The mechanism established a base dividend rate of 10%. If gas prices increased above a base level the dividend rate was reduced according to a sharing formula. However, if gas prices fell below the base level the dividend rate did not increase (a "one-way" sliding scale). The mechanism was made symmetric in 1867. Note that the mechanism was not mandatory and it was introduced during a period of falling prices [Hammond, Johnes, and Robinson (2002, pp. 255–256)]. A related profit sharing mechanism [what Hammond, Johnes, and Robinson (2002) call the "Basic Price System"] was introduced in 1920 that provided a minimum guaranteed 5% dividend to the firm's shareholders and shared changes in revenues from a base level between the consumers, the owners of the firm and the firm's employees. Specifically, this mechanism established a basic price p_b to yield a 5% dividend rate. This dividend rate was the minimum guaranteed to the firm. At the end of each financial year the firm's actual revenues (R) were compared to its basic revenues $R_b = p_b$ times the quantity sold. The difference between R and R_b was then shared between consumers, investors and employees, apparently subject to the constraint that the dividend rate would not fall below 5%. Hammond, Johnes, and Robinson (2002) use "data envelopment" or "cost frontier" techniques [Giannakis, Jamasb, and Pollitt (2004)] to evaluate the efficiency properties of three alternative gas distribution pricing mechanisms used in England based on data for 1937. While they find significant differences in performance associated with the different mechanisms, the linkage between the incentive structure of the different mechanisms and the observed performance is unclear. Moreover, the analysis does not appear to account for the potential endogeneity of the choice of regulatory mechanism applied to different firms.

In the early 20th century, economists took note of the experience with sliding scale mechanisms in England, but appear to have concluded that they were not well matched to the regulation of electricity and telephone service (and other sectors) where demand and technology were changing fast and future costs were very uncertain [Clark (1913)]. As already discussed, cost of service regulation (with regulatory lag and prudence reviews) evolved as the favored alternative in the U.S., Canada, Spain and other countries

[24] This is before the development of natural gas. "City gas" was manufactured from coal by local gas distribution companies. At the time there were both private and municipal gas distribution companies in operation in England.

with private (rather than state-owned) regulated monopolies and the experience in England during the 19th and early 20th centuries was largely forgotten by both regulators and students of regulation.

State public utility commissions began to experiment with performance based regulation of electric utilities in the 1980s. The early programs were targeted at specific components of an electric utility's costs or operating performance such as generation plant availability, heat rates, or construction costs [Joskow and Schmalensee (1986), Sappington et al. (2001)]. However, formal comprehensive incentive regulation mechanism have been slow to spread in the U.S. electric power industry [Sappington et al. (2001)], though rate freezes, rate case moratoria, price cap mechanisms and other alternative mechanisms have been adopted in many states, sometimes informally since the mid-1990s. Because of the diversity of these programs, the co-existence of formal and informal programs, and the simultaneous introduction of wholesale and retail competition and related vertical and horizontal restructuring initiatives (Joskow, 2000), it has been difficult to evaluate the impact of the introduction of these incentive regulation mechanisms in the electric power sector. Rose, Markowicz, and Wolfram (2004) examine aspects of the operating performance of regulated generating plants during the period 1981–1999 and find that the threat of the introduction of retail competition led to improvements in various indicia of generating plant performance.

Beginning in the mid-1980s a particular form of incentive regulation was introduced for the regulated segments of the privatized electric gas, telephone and water utilities in the U.K., New Zealand, Australia, and portions of Latin American as well as in the regulated segments of the telecommunications industry in the U.S.[25] The mechanism chosen was the "price cap" [Beesley and Littlechild (1989), Brennan (1989), Armstrong, Cowan, and Vickers (1994), Isaac (1991), Joskow (2006)]. In theory, a price cap mechanism is a high-powered "fixed price" regulatory contract which provides powerful incentives for the firm to reduce costs. Moreover, if the price cap mechanism is applied to a (properly) weighted average of the revenues the firm earns from each product it supplies, the firm has an incentive to set the second-best prices for each service [Laffont and Tirole (2000), Armstrong and Vickers (1991)].

In practice, price cap mechanisms apply elements of cost of service regulation, yardstick competition, high powered "fixed price" incentives plus a ratchet. Moreover, the regulated firm's ability to determine the structure of prices under an overall revenue cap is typically limited. Under price cap regulation the regulator sets an initial price p_o (or a vector of prices for multiple products). This price (or a weighted average of the prices allowed for multiple products) is then adjusted from one year to the next for changes in inflation (rate of input price increase or RPI) and a target productivity change factor "X." Accordingly, the price in period 1 is given by:

$$p_1 = p_o(1 + \text{RPI} - X) \tag{50}$$

[25] The U.S. is behind many other countries in the application of incentive regulation principles, though their use is spreading in the U.S. beyond telecommunications.

Typically, some form of cost-based regulation is used to set p_o. The price cap mechanism then operates for a pre-established time period (e.g. 5 years). At the end of this period a new starting price and a new X factor are established after another cost-of-service and prudence or efficiency review of the firm's costs. That is, there is a pre-scheduled regulatory-ratchet built into the system.

Several things are worth noting about price cap mechanisms since they have become so popular in the regulatory policy arena. A pure price cap without cost-sharing (a sliding scale mechanism) is not likely to be optimal given asymmetric information and uncertainty about future productivity opportunities. Prices would have to be set too high to satisfy the firm participation constraint and too much rent with be left on the table for the firm. The application of a ratchet from time to time that resets prices to reflect observed costs is a form of cost-contingent dynamic regulatory contract. It softens cost-reducing incentives but extracts more rents for consumers.

Although it is not discussed too much in the empirical literature, price cap mechanisms are typically focused on operating costs only, with capital cost allowances established through more traditional utility planning cost-of-service regulatory methods. In addition, it is widely recognized that a pure price cap mechanism provides incentives to reduce both costs and the quality of service. Accordingly, price cap mechanisms are increasingly accompanied either by specific performance standards and the threat of regulatory penalties if they are not met or formal PBR mechanisms that set performance standards and specify penalties and rewards for the firm for falling above or below these performance norms [OFGEM (2004b, 2004c, 2004d), Sappington (2003), Ai and Sappington (2002), Ai, Martinez, and Sappington (2004)].

A natural question to ask about price cap mechanisms is where does "X" (and perhaps p_o) come from [Bernstein and Sappington (1999)]? Conceptually, assuming that RPI is a measure of a general input price inflation index, X should reflect the difference between the expected or target rate of total factor productivity growth for the regulated firm and the corresponding productivity growth rate for the economy as a whole and the difference between the rate of change in the regulated firm's input prices and input prices faced by firms generally in the economy. That is, the regulated firm's prices should rise at a rate that reflects the general rate of inflation in input prices less an offset for higher (or lower) than average productivity growth and an offset for lower (or higher) input price inflation. However, the articulation of this conceptual rule still begs the question of how to calculate X in practice.

In practice, the computation of X has often been fairly ad hoc. The initial application of the price cap mechanism by the Federal Communications Commission (FCC) to AT&T's intercity and information services used historical productivity growth and added an arbitrary "customer dividend" to choose an X that was larger than the historical rate of productivity growth. In England and Wales and some other countries, benchmarking methods have come to be used to help to determine a value for X [Jamasb and Pollitt (2001, 2003)] in a fashion that is effectively an application of yardstick regulation. A variety of empirical methods have been applied to identify a cost efficiency frontier and to measure how far from that frontier individual regulated firms

lie. The value for X is then defined in such a way as to move the firms to the frontier over a pre-specified period of time (e.g. five years). These methods have recently been expanded to include quality of service considerations [Giannakis, Jamasb, and Pollitt (2004)].

The extensive use of periodic "ratchets" or "resets to cost" along with price cap mechanisms reflect the difficulties of defining an ideal long-term value for X and the standard tradeoffs between efficiency incentives, rent extraction and firm viability constraints. These ratchets necessarily dull incentives for cost reduction. Note in particular that with a pre-defined five year ratchet, a dollar of cost reduction in year one is worth a lot more than a dollar of cost reduction in year four since the cost savings are retained by the firm only until the next reset anniversary [OFGEM (2004b)].

Most of the scholarly research evaluating the effects of incentive regulation have focused on the telecommunications industry [Kridel, Sappington, and Weisman (1996), Tardiff and Taylor (1993), Crandall and Waverman (1995), Braeutigam, Magura, and Panzar (1997), Ai and Sappington (2002), Banerjee (2003)]. Ai and Sappington's study is the most recent and comprehensive. They examine the impact of state incentive regulation mechanism applied to local telephone companies between 1986 and 1999 on variables measuring network modernization, aggregate investment, revenue, cost, profit, and local service prices. The methodological approach involves the use of a panel of state-level observations on these performance indicia, state regulatory regime variables and other explanatory variables. Instrumental variables are used to deal with the endogeneity of the choice of regulatory regime and certain other explanatory variables so that the fixed-effects estimates are consistent. Ai and Sappington (2002) find that there is greater network modernization under price cap regulation, earnings sharing regulation, and rate case moratoria (effectively price cap regulation with $\text{RPI} + X = 0$), than under rate of return regulation. Variations in regulatory mechanisms have no significant effects on revenue, profit, aggregate investment, and residential prices, and except for rate case moratoria, on costs. Crandall and Waverman (1995) find lower residential and business prices under price cap regulation than under rate of return regulation but other forms of incentive regulation do not yield lower prices. Tardiff and Taylor (1993) use similar methods and find similar results to Ai and Sappington (2002). Braeutigam, Magura, and Panzar (1997) find lower prices under some types of price cap regulation but not under other form of incentive regulation.

Sappington (2003) reviews several studies that examine the effects of incentive regulation on the quality of retail telecommunications service in the U.S. These studies do not lead to consistent results and for many dimensions of service quality there is no significant effect of variations in the regulatory regime applied. Ai, Martinez, and Sappington (2004) also examine the effects of incentive regulation on service quality for a state-level panel covering the period 1991 through 2002. They find that incentive regulation is associated with significantly higher levels of service quality compared to rate of return regulation for some dimensions of service quality (e.g. installation lags, trouble reports, customer satisfaction) and significantly lower levels of service quality in other dimensions (e.g. delays in resolving trouble reports, percentage of installation

commitments met). Banerjee (2003) provides a related empirical analysis of the effects of incentive regulation on telephone service quality.

Systematic research on the effects of incentive regulation in other industries is limited. Newbery and Pollitt (1997) argue that the incentive regulatory mechanisms applied to electricity distribution companies in England and Wales during the first half of the 1990s led to significant efficiency improvements. Significant savings associated with the application of price cap and other incentive mechanisms to electricity distribution and transmission have also been noted by regulators in the U.K. [OFGEM (2004a, 2004b), Joskow (2006)]. Rudnick and Zolezzi (2001) examine the changes in several dimensions of productivity in the liberalized electricity sectors in Latin America during the 1990s and find significant improvements in these productivity indicia. Bacon and Besant-Jones (2000) provide a broader assessment of the effects of privatization, market liberalization and regulatory reform of electricity sectors in developing countries, indicating more mixed results. However, it is hard to know how much of these observed cost reductions is due to the incentive regulation mechanisms and how much to privatization. Estache and Kouasi (2002) examine the diverse performance effects of alternative governance arrangements on African water utilities and Estache, Guasch, and Trujillo (2003) analyze the effects of price caps and related incentive provisions on and the renegotiation of infrastructure contracts in Latin America.

Gagnepain and Ivaldi (2002) examine the effects of incentive regulatory policies on public transit systems in France using data on a panel of French municipalities over during the period 1985–1993. This is a particularly interesting paper because the empirical analysis is embedded directly in a structural model of optimal regulation a la Laffont and Tirole (1993) discussed above.

Since 1982 local public transport (buses, trams) in France has been decentralized to the municipalities. The municipalities own the rolling stock and infrastructure but contract out the operation of the systems to private operators in 80% of the cases (there are three private operators in the country which also provide other municipal services). Fares (P) do not produce enough revenue to cover the total costs incurred by the operators so there are transfer payments from the government to the operator to satisfy break-even constraints (the treatment of the costs of municipal-owned infrastructure in the analysis is a little unclear, but they are not paid directly by the operator).

Private operators are given either "cost-plus" (CP) contracts or "fixed price" (FP) contracts. The former cover observed costs and ex post deficits. The latter cover expected costs and expected deficits. In 1995, 62% of the operators had fixed price contracts and 25% cost-plus contracts. The rest were operated by the municipalities or are not in the data base for other reasons. The contracts have a duration of one year and municipalities apparently never switch operators during the sample period. The analysis focuses on the larger municipalities with more than 100,000 population (excluding Paris, Lyon, Marseilles, which were not included in the data set) and relies on a panel data set for 59 municipalities over 1985–1993 period on input costs, output, network infrastructure.

The paper includes the following empirical analyses: (a) estimates the parameters of a cost function (structural model) for urban public transport that treats the effects

of regulation on costs under asymmetric information as endogenous given the type of contract each system is place upon; (b) estimates the parameters of the distribution of the "labor inefficiency" parameter θ and the cost of effort function give assumptions about the form of these functions (e.g. a beta distribution for θ) and the cost function (Cobb-Douglas technology); (c) estimates the level of inefficiency θ_i and effort (e_i) of each urban transport system given the cost function's parameters, the estimated parameter of the cost of effort function and the regulatory contract they have been placed upon; (d) estimates the implied cost of public funds given the cost function parameters, the parameters of a demand function for public transport and the form of the contract each transport operator has been given assuming that the municipality sets the optimal fare (Ramsey-Boiteux) given demand, costs, and each municipality's cost of public funds λ and; (e) calculates the optimal regulatory contract [second-best under asymmetric information a la Laffont and Tirole (1993)] for each system given the cost of public funds and its inefficiency parameter and the welfare gains from doing so.

These analyses leads to several conclusions including: (a) there are economies of scale in urban transport; (b) there is a large variation in the efficiency parameters for different networks; (c) for the lowest θ group the cost distortions (difference in efficiency between a fixed price and a cost plus contract) are not significantly different between the FP and CP contracts, for the intermediate θ the difference in cost distortions is about 4%, and for the highest θ group there is a lot of inefficiency even with a FP contract, but FP contracts reduce costs significantly (mix of CP and FP contracts); (d) cost of public funds varies from 0.17 to 0.56 across municipalities and (e) optimal second best (Laffont-Tirole) contracts improve welfare significantly compared to cost plus contracts, but not compared to fixed price contracts.

Research measuring the effects of regulation and deregulation on the speed of introduction of new services and technologies in telecommunications also make it clear that dynamic considerations are extremely important from a social welfare perspective [Hausman (1997), Crandall and Hausman (2000)]. Hausman (1997) estimates the costs of FCC delays in the introduction of voice messaging service and cellular telephone service by estimating the structural parameters of consumer demand and the value to consumers of new goods. The basic method is to estimate the effect of the introduction of a new good on real consumer income and then to perform a counterfactual analysis to measure the costs foregone by regulatory delays in introducing the product. He finds that regulatory and court delays led voice messaging to be introduced 5 to 7 years later than it would have been without these delays. He finds as well [see also Hausman (2002, 2003)] that FCC regulatory delays led to cellular telephone being introduced 7 to 10 years later than would have been the case without these delays. The social costs of these delays are estimated to be about $6 billion and $30 billion in 1994 dollars respectively. Hausman (2002) also finds that other regulatory restrictions on mobile service competition led to significantly higher prices for mobile services.

Greenstein, McMaster, and Spiller (1995) examine the effects of incentive regulation on the deployment of digital technologies by local telephone carriers. Recent work by Thomas Hubbard has shown how new technologies adopted by post-deregulation

trucking firms have both served to improve service quality and to improve productivity and lower costs [Hubbard (2001, 2003)]. Regulation of prices and entry prior to deregulation in 1980 inhibited the diffusion of these kinds of technologies in a number of different ways, though the precise impact of regulation per se has not been measured. Rose and Joskow (1990) examine the diffusion of new electric generating technologies in the electric power sector. Goolsbee and Petrin (2004) find significant consumer benefits from the entry of direct broadcast satellite to compete with cable TV, but limited effects on the cable firms' marker power.

It is clear that the social costs of delaying product and process innovations can be very significant. Both theoretical and empirical research has probably focused too much on static welfare effects associated with the impacts of regulation on prices and costs in the short run and too little research has focused on the effects of regulation on the adoption and diffusion of product and process innovations.

10. Competitive entry and access pricing

The firms in many industries that have been subject to price and entry regulation have organizational structures that involve vertical integration between production of complementary services at different levels of the production chain. For example, in most countries, electric power companies historically evolved with governance structures where generation, transmission, distribution and retail marketing of electricity where vertically integrated (Joskow, 1997). However, there are also thousands of small municipal and cooperative distribution utilities that purchase power from third parties (typically proximate vertically integrated utilities) which they then resell to retail consumers to whom they provide distribution (delivery) service in their franchise areas. In many countries natural gas producers also own natural gas pipelines that transport natural gas from where it is produced to where it is distributed in local consumption areas. Telephone companies historically provided both local and intercity services and, in the U.S., the vertical integration extended into the production of telephone network equipment and customer premises equipment.

These industries likely evolved with these structures in response to economies of vertical integration [Joskow (1997)]. However, to the extent that the economies of vertical integration led to the integration of a production segment with natural monopoly characteristics with a production segment without natural monopoly characteristics, the effect of vertical integration is to extend the natural monopoly to the potentially competitive segments as well. For example, the transmission and distribution of electricity have natural monopoly characteristics. However, there are numerous generating plants in each region of the U.S., suggesting that the generation of power may be potentially competitive [Joskow and Schmalensee (1983), Joskow (1997)]. Vertical integration effectively extends the natural monopoly over transmission and distribution to generation when firms in the industry are vertically integrated, extending the boundaries of regulation and its complexities and potential imperfections. Alternatively, two or more vertically integrated segments may once have had natural monopoly characteristics as well as

economies of vertical integration, but technological change may have changed the characteristics of the underlying technology at one or more levels of the vertical chain to make it potentially competitive. For example, microwave, satellite, and radio technology, as well as the diffusion of cable television, have changed the economic attributes of both the supply and demand for intercity and local telecommunications services dramatically.

The bundling of multiple supply segments (or products), one or more of which does not have natural monopoly characteristics and is potentially competitive, into a single firm subject to price and entry regulation naturally leads to a number of questions and issues. Would better performance be achieved by separating the potentially competitive segments from the natural monopoly segments and removing price and entry regulation from the competitive segments? Are the benefits of potentially imperfect competition in these segments greater than the lost cost savings from vertical integration (if any)? Or should we allow the incumbent regulated firm to continue to offer both sets of services, but allow competitive entry into the potentially competitive segments so that entering firms can compete with the incumbent? If we take this approach when and how do we regulate and deregulate the prices charged by the incumbent for competitive services? How do we know that competitive entry will take place because lower cost suppliers have incentives to enter the market rather than inefficient entry resulting from price distortions resulting from decades of regulation? Should limits be placed on the ability of regulated firms to respond to competitive entry to guard against predatory behavior? Is structural separation necessary (divestiture) or is functional separation with line of business restrictions to deal with potential cross-subsidization of by regulated services by regulated services and behavior that disadvantages competitors sufficient to foster efficient competition? These issues are especially challenging in many regulated industries because access to the natural monopoly segments (e.g. the electric transmission network) is necessary for suppliers in the competitive segment (e.g. generating plants) to compete. Such networks are often referred to as "essential facilities" or "bottleneck facilities," though these terms have been abused in the antitrust policy context.

The terms and conditions under which competitive suppliers can gain access to the incumbent's monopoly network when the incumbent is also a competitor in the competitive segments has been the focus of considerable research in the last decade as previously regulated vertically integrated firms in several regulated industries are "restructured" to separate natural monopoly network segments from competitive segments and price and entry regulation relaxed in the competitive segments [Vickers (1995), Laffont and Tirole (2000), Baumol and Sidak (1994), Vogelsang (2003)]. If the access prices are set too low, inefficient entry may be encouraged. If access prices are set too high they will serve as a barrier to entry to competitors who are more efficient than the entrant or encourage inefficient bypass of the network to which access is sought. When prices for regulated services are partially based on realized costs, cost allocations between regulated and unregulated services becomes an issue as well since the incumbent may be able to subsidize the costs of providing competitive services by hiding some of them in the cost of service used for determining regulated prices. Access pricing issues

also arise when the incumbent network operator's business is restricted to regulated network services only, but the nature of the distortions is different as long as all competitors are treated equally.

10.1. One-way network access

Much of the access pricing literature initially evolved in the context of the development of competition in the supply of intercity communications services and the interconnection of competing intercity networks with regulated monopoly local telephone networks which originate and terminate intercity calls. I will focus on telecommunications examples here, following the development in Laffont and Tirole (2000). Conceptually similar issues arise in electricity and natural gas as well, though the technical details are different (Joskow, 2005a). There are two kinds of services. The first is provision of "local network" service which is assumed for now to be a natural monopoly and subject to price and entry regulation. The second service is intercity service which allows for transmission of voice and data signals between local networks in different cities, is supplied by the incumbent and is being opened to potential competitors. The incumbent is assumed to be vertically integrated into both the provision of local exchange services and the provision of intercity services and the prices for both services are assumed to be regulated. For a competitive intercity supplier to enter the market and compete with the incumbent it must be able to gain access to the local network in one city to originate calls and to gain access the local network in the other city to complete the calls. The entrant is assumed to provide its own intercity facilities to transport the calls between local networks but relies on the regulated monopoly incumbent to provide local connection services on the local origination and termination networks. These relationships are displayed in Figure 12.

Let:

q_o = quantity of local calls sold at price p_o

q_1 = quantity of incumbent's long distance calls at price p_1

q_2 = quantity of entrant's long distance calls at price p_2

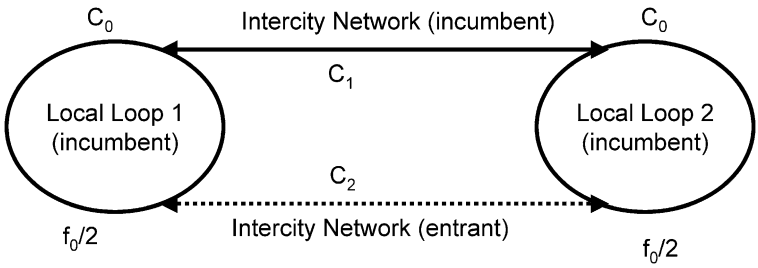


Figure 12. One-way access.

$Q = q_o + q_1 + q_2 =$ total calls

$f_o =$ fixed cost of the local network

$c_o =$ cost of originating or terminating a local call

$c_1 =$ incumbent's cost of a long distance call

$c_2 =$ entrant's cost of a long distance call

Convention: Every local or long distance call involves one origination and one termination on the local network. Each local network has a marginal cost per call of c_o so the marginal cost to use local networks to originate and complete a call is $2c_o$.

Incumbent's costs: $f_o + 2c_o(q_o + q_1 + q_2) + c_1 q_1$

Entrant's costs: $c_2 q_2 + aq_2 = c_2 q_2 + (p_2 - c_2)q_2$

where "a" is the *access price* the entrant must pay to the incumbent for using its local network facilities (one origination and one termination per long distance call).

Assume that the entrant has no market power so it sets its price for intercity calls equal to the marginal cost it incurs, including the price it is charged for access to the incumbent's local access. The entrant's price for long distance service p_2 must be (it passes along marginal costs with no additional markup):

$$p_2 = a + c_2$$

and

$$a = p_2 - c_2$$

(A useful way to think about this is that the incumbent "subcontracts" with the entrant to supply the entrant's long distance service at cost.)

The incumbent's profits on the provision of local, long distance and "access service" are then given by:

$$\begin{aligned}\pi(p_o, p_1, p_2) = {} & (p_o - 2c_o)q_o \\ & + (p_1 - c_1 - 2c_o)q_1 \\ & + (p_2 - c_2 - 2c_o)q_2 \\ & - f_o\end{aligned}$$

(where $p_2 - c_2 = a$).

Assume that $S_o(p_o)$ and $S_1(p_1, p_2)$ give the net consumers' surpluses for local and long distance and recall that the derivative of the net surplus with respect to a price is (minus) the corresponding quantity. Assume as well that the incumbent is regulated and is subject to a breakeven constraint. Then the optimal prices p_o, p_1, and p_2 (and "a") are given by:

$$\underset{(p_o, p_1, p_2)}{\text{Max}} \left\{ S_o(p_o) + S_1(p_1, p_2) + \pi(p_o, p_1, p_2) \right\} \tag{51}$$

$$\text{S.T. } \pi(p_o, p_1, p_2) \geq 0$$

This is just a familiar Ramsey-Boiteux pricing problem and yields the following familiar conditions where λ (> 0) is the shadow cost of the budget constraint and the η_i are price *superelasticities* that account for cross-price effects when there are goods that are substitutes or complements:

$$\frac{p_0 - 2C_0}{p_0} = \frac{\lambda}{1+\lambda}\frac{1}{\eta_0} \tag{52}$$

$$\frac{p_1 - C_1 - 2C_0}{p_1} = \frac{\lambda}{1+\lambda}\frac{1}{\eta_1} \tag{53}$$

$$\frac{p_2 - C_2 - 2C_0}{p_2} = \frac{\lambda}{1+\lambda}\frac{1}{\eta_2} \tag{54}$$

Note that if there were no fixed costs f_o, the optimal price would be equal to marginal cost and the access price "a" would equal $2c_o$. However, with fixed costs, the access price includes a contribution to these fixed costs. The optimal Ramsey-Boiteux access price $a = p_2 - c_2$ [Laffont and Tirole (1996, 2000, pp. 102–103)] then follows from the formula

$$a = 2C_0 + \lambda/(1+\lambda)(p_2/\eta_2)$$

This can be rewritten (Armstrong, Doyle, and Vickers, 1996) as

$$a = 2C_0 + \lambda/(1+\lambda)(p_2/\varepsilon_2) + \delta(p_1 - 2C_0 - C_1) \tag{55}$$

where

$$\delta = -\left[\frac{\partial q_1/\partial p_2}{\partial q_2/\partial p_2}\right]$$

is the change in the sales of the incumbent divided by the change in sales of the competitive entrant and ε_2 is the own-price elasticity of demand without accounting for cross-price effects. The Ramsey-Boiteux access price "a" is set above marginal cost and therefore contributes to the incumbent's fixed costs. It is composed of two components. The first is the standard Ramsey price equation. The second allows for the substitution between the incumbents sales of network services and its loss of retail sales to the competitive entrant.

According to Willig (1979) the appropriate access price is given by $a = p_1 - c_1$ or the difference between the incumbent's retail price for long distance calls and the marginal or avoided cost of supplying these calls (see also Baumol and Sidak, 1994, Baumol, Ordover, and Willig, 1997). This rule is often referred to as the Efficient Component Pricing Rule or ECPR. It has been argued that this rule has a number of desirable features including: (a) potential entrants can enter profitably if and only if they are more efficient than the incumbent; since $a + c_2 = p_1 - (c_1 - c_2)$, only a cost advantage will lead to entry; (b) entry is neutral regarding operating profit for the incumbent since it still gets the same profits on sales of "access" as it does on retail sales. Incumbent does

not have incentive to destroy entrant; (c) entry does not interfere with existing cross-subsidies and is not "unfair" to the incumbent; (d) if entrants do not have lower costs there will be no entry and (e) if entrants have lower costs the incumbent will be driven from the retail market and will supply only "access."

Laffont and Tirole (1996, 2000) and others point out a number of problems with the ECPR. They include: (a) ECPR is a "partial rule" in the sense that it does not tell us how p_1 should be set optimally. It takes p_1 as given. However, given that regulators are unlikely to have set second-best efficient prices as competition emerges, this may be a very practical real world approach; (b) ECPR is implied by Ramsey-Boiteux pricing only under a restrictive set of assumptions. These assumptions are equivalent to assuming that there is full symmetry between the incumbent and the potential entrant in the sense that they have equal costs of providing intercity services ($c_1 = c_2$), that they face symmetrical demands in the intercity (competitive) segment, and that the entrants have no market power. In this case since $a = p_2 - c_2$, the combination of $p_1 = p_2$ and $c_1 = c_2$ implies that $a = p_1 - c_1$ which is the efficient component pricing rule; (c) ECPR gives the wrong access price for competitive services that are differentiated products rather than being identical to the product produced by the incumbent.

To illustrate this last point, assume that the competitors have the same costs by different demands

$$q_1 = a_1 - bp_1 + dp_2$$
$$q_2 = a_2 - bp_2 + dp_1$$

where $a_1 > a_2$ (brand loyalty/less elastic demand) and $b > d$.

In this case is can be shown that the access price should be lower than ECPR:

$$a < p_1 - c_1 \quad \text{and} \quad p_1 > p_2$$

The reason is that the optimal price for the incumbent is higher than the optimal price for the entrant because the incumbent has a less elastic demand:

$$p_1 > p_2$$

The access price (for an intermediate good) must be lower to keep p_2 from rising above its optimal level. In this case, if the incumbent has lower costs than the entrant the access price should be higher than ECPR. The logic is the same in reverse. The optimal prices are $p_1 < p_2$.

The ECPR is also not efficient if entrants have market power. If the entrants have market power they will mark up the access price when they set the retail price leading to a classic double marginalization problem [Tirole (1988, Chapter 4)]. When there is competitive entrant with market power, the optimal access prices is the ECPR level minus the competitor's unit markup m:

$$a = p_1 - c_1 - m$$

Finally, the entrant may be able and willing inefficiently to bypass the incumbent's network if the access price is greater than its own cost of duplicating the network.

The regulator can respond to this problem by setting access charge lower than the incumbent's entry cost. But this increases the incumbent's access deficit and the incumbent would than have to increase prices further for captive customers. In principle the regulator could charge low access price and then levy an effective excise tax on the competitor's sales to cover the access deficit, but such an instrument may not be available to the regulator.

These considerations suggest that setting the optimal access price requires consideration of many other aspects of the industrial organization of the potentially competitive sector which may not be consistent with the assumptions that lead to the ECPR. The optimal access prices reflect standard Ramsey pricing considerations, the relationship between wholesale access prices and the incumbent's retail sales, differentiated product and double marginalization (vertical market power) considerations, and imperfect competition in the potentially competitive segment. Setting the optimal access prices clearly places very significant information and computational burdens on regulators.

Laffont and Tirole (1996, 2000) suggest that a superior approach to setting access prices is to apply a *global price cap* to the regulated firm that includes "network access" as one of the products included in the price cap. If the weights in the price cap formula are set "properly" (equal to the realized quantities from optimal pricing) then the regulated firm will have the incentive to price all of the services covered, including pricing "access" to the network at the optimal Ramsey-Boiteux prices that take all of the relevant costs and super-elasticities into account. They recognize, however, that finding the optimal quantities also creates a significant information and computational burden on regulators. In addition, applying a price cap mechanism in this way may enhance incentives for the incumbent to adopt a predatory pricing strategy that leads to an access price that is set too low, with the lost net revenue partially recouped in the short run by increasing prices for other regulated services and in the long run by inducing competitors to exit the market. Accordingly, they suggest that a global price cap be combined with a rule that the access price can be no lower than the difference between the incumbent's retail price and its avoided cost or its "opportunity cost" of sales lost to the incumbent ($a \leq p_1 - c_1$).

10.2. Introducing local network competition

In 1996, the U.S. Congress determined that competition should be opened up for providing local network services as well as intercity services. That is, it adopted a set of policies that allowed competitors to offer local telephone services in competition with the incumbent Local Exchange Carriers (ILECs). The Competitive Local Exchange Carriers (CLECs) could compete by building their own facilities (as a cable television network might be able to do) or by leasing the facilities owned by the ILEC. The argument was that while there were opportunities for facilities-based competition at the local network level, there were likely to be components of the local network that had natural monopoly characteristics (e.g. the "last mile" from the local exchange to the end-user's premises). At the very least, it would take time for facilities based competitors to build

out a complete network. This is not the place to go into the complex issues associated with local service competition, but this policy required regulators to set regulated prices at which competitors could gain access to the local loop. Accordingly, a brief discussion of the issues associated with the regulated pricing of "unbundled network elements" is in order (Laffont and Tirole, 2000, Crandall and Hausman, 2000).

Following the Telecommunications Act of 1996, the FCC required ILECs to offer to lease pieces of their networks (network elements) to CLECs. In addition to requiring interconnection of networks so that all termination locations could be reached on any network, the FCC concluded that it would also require ILECs to lease individual network elements to CLECs. The FCC decomposed ILEC networks into a complete set of "Network Elements" and required the ILECs to lease each and every element to CLECs requesting service "at cost" (see Hausman, 1999, Crandall and Hausman, 2000). For example, RCN built its own network providing cable TV, telephone and high-speed internet service in portions of Boston and Brookline. It is interconnected to Verizon's local telephone network so that RCN subscribers can reach on-net and off-net locations and vice versa. If RCN also wanted to offer service to potential subscribers in, say, Cambridge, but building its own network there is uneconomical, it could then lease all of the network elements on Verizon's Cambridge's network at "cost-based" wholesale prices, and begin offering service there as if it were its own network. At that time, the FCC thought that this was the best way to promote local service competition. This policy leads to a number of questions, only two of which I will discuss briefly here.[26]

What is the right regulated price for network elements? The Federal Communications Commission (FCC) used an engineering model of a local telephone network and estimates of the current cost of equipment and maintenance to build an "optimal network" and then to estimate the "forward looking long run incremental costs" for each network element (TELRIC). This approach has a number of shortcomings: (a) The underlying engineering model is at best an imperfect representation of real telephone networks; (b) the cost calculations fail properly to take into account economic depreciation of equipment and lead to current cost estimates that are biased downward [see Hausman (1999, 2003)]; (c) the cost calculations fail to take into account the interaction between the sunk cost nature of telecom network investments and uncertainty over future demand, equipment prices, and technical change. This also leads to a significant underestimate of the true economic costs of short-term leasing arrangements. The FCC leasing rule effectively includes an imbedded option. The CLEC can take the service for a year and then abandon it if a cheaper alternative emerges or continue buying at wholesale until a better alternative does emerge (if the ILECs instead could sell the

[26] Other questions include: If RCN is simply buying service on Verizon's network at wholesale prices (including connects and disconnects, network maintenance, etc.) and then reselling these services under it's brand name, what is the social value added from making this competition possible? "Retail service" costs are very small. If an ILEC must lease any and all of its facilities to competitors "at cost," how does this affect its incentives to invest on its network in general, and in particular, to invest in new technologies for which it must compete with other firms?

network elements to the CLECs at their installation cost rather than offer the service on short-term leases, this would solve this problem) [Hausman (1999), Hausman and Myers (2002), Pindyck (2004)]; (d) wholesale network element prices determined by the TELRIC rules are substantially below the ILECs actually regulated costs. This is not surprising since the regulated costs are based on traditional "depreciated original cost ratemaking" techniques and reflect historical investments that were depreciated too slowly. This creates stranded cost problems for the ILECs and potential distortions in demand for "new network elements" rather than equivalent "old network elements." Unlike the situation in electricity and natural gas sector reforms, where regulators have recognized and made provisions for stranded costs recovery, this issues has largely been ignored in the U.S. in the case of telecom reform; and (e) these rules reduce ILEC incentives to invest in uncertain product service innovations. The FCC rules ignore the cost of "dry holes." CLECs can buy new successful services "at cost," compete with the ILEC for customers for these services, and avoid paying anything for ILEC investments that are unsuccessful [Crandall and Hausman (2000), Pindyck (2004)].

All of these considerations suggest that TELRIC underprices network elements. Moreover, competitive strategies of CLECS may be driven more by imperfections in FCC pricing rules than by their ability to offer cheaper/better products. Nevertheless, so far there has been only limited successful CLEC competition except for business customers in central cities.

10.3. Two-way access issues

Opening up the local loop to competition raises another set of interesting pricing issues that arise when there are two (or more) bottleneck networks which (a) need to interconnect (cooperate) with one another to provide retail services and (b) may compete with one another for customers. Such situations include (a) overlapping LECs which require interconnection; (b) internet networks which exchange data traffic; (c) credit card associations, and (d) international telephone calls where networks at each end must exchange traffic. These situations create a set of "two-way access" pricing problems. What are the most efficient access pricing arrangements to support interconnection between the two (or more) networks and what institutional arrangements should be relied upon to determine threes prices? Regulation, cooperation, non-cooperate competitive price setting? There are a number of policy concerns: (a) cooperation may lead to collusion to raise prices at the level where the networks compete (e.g. retail calling); (b) non-cooperative access pricing may fail to account properly for impacts on other networks; and (c) inefficient access prices may increase entry barriers and soften competition [Laffont and Tirole (2000); Laffont, Rey, and Tirole (1998a, 1998b)].

The literature on two-way access pricing is closely related to the growing and much broader literature on "two-sided markets" [Rochet and Tirole (2003, 2004)], though he precise definition of what is included within the category "two-sided markets" is somewhat ambiguous. However, many markets where network platforms are characterized by network externalities are "two-sided" in the sense that the value of the network plat-

forms depends on getting buyers and/or sellers on both sides of the market to use them effectively through pricing arrangements and market rules. The value of a credit card to consumers depends on its broad acceptance by retailers. The value of a telephone network depends on the number of consumers who can be reached (to call or be called) on it or through interconnection with other telephone networks. The value of a bank's ATM network to its depositors depends on their ability to use other ATM networks to get cash from their bank accounts.

A discussion of the literature on two-sided markets in beyond the scope of this chapter. However, to identify the issues at stake I will briefly discuss the nature of the access pricing issues that arise absent price regulation when multiple networks serve consumers who in turn value reaching or being reached by consumers connected to the other networks. While these kinds of problems may be solved by regulation, the more typical solution is for the network participants and the networks to negotiate access pricing arrangements and market rules to deal with the potential inefficiencies created by network externalities and market power. I will follow Laffont and Tirole (2000) to identify some of the issues at stake in this literature.

Consider a situation where we have a city served by two local telephone networks which we can call the A and B networks (Weiman and Levin, 1994). Customers are connected to one network or the other and all customers need to be able to call any other customer whether they are on the same network or not. Laffont and Tirole's (2000) analysis of this situation adopts two "conventions." (a) The calling company's network pays a (per minute) termination (access) charge "a" to the termination company's network and can bill the caller for this charge. The receiving customer does not pay a termination charge for the call (this is known as a "caller pays" system; and (b) retail prices are unregulated so that networks are free to charge whatever they conclude is profit maximizing for sales to final consumers. The question is whether competition is likely to lead to efficient outcomes absent any regulatory rules.

Assume that there are 100 consumers each connected to a separate independent network who call each other. Each network sets it own access charge for terminating calls to it. The originating network incurs marginal cost c to get the call to the network interface and then a termination charge a_i to the receiving network. Assume that each originating network sets a retail price equal to c plus the average of the termination charges of the 99 networks to which it interconnects (no price differences based on the location of the termination network). In this case, the impact of an increase in a_i on the average termination price originating callers pay to call network i is very small. In this case each network has an incentive to charge high access charges because the perceived impact on the volume of calls that it will receive is small. All networks set access fees too high and the average access fee passed along to consumers by the calling networks is too high. This in turn leads to high retail prices and too little calling.

This result is most striking for a large number of networks and no network specific price discrimination. However, Laffont and Tirole (2000) show that similar results emerge when there are only two networks which have market power. Consider the case of international calls. There are two monopolies (one in each country) and each

sets a termination charge that applies to calls received from the other. Since each is a monopoly whatever the access charge is chosen by the other, this will get "marked up" by the local monopoly leading to a double marginalization problem [Laffont and Tirole, (2000, Box 5.1)].

It should be obvious as well that if two networks which compete intensely (Bertrand) with one another at retail *cooperated* in setting their respective access prices that they could agree on high access prices, increasing the perceived marginal cost at retail and the associated retail prices. The monopoly profits would then reside at the wholesale (access business) level rather than the retail level. Basically, it is profitable for each network to increase its rivals costs so that market prices rise more than do the firm's costs (Ordover, Saloner, and Salop, 1990). Accordingly, both non-cooperative and co-operative access pricing can lead to excessive retail prices. In the context of a simple duopoly model with competing firms selling differentiated products, Laffont and Tirole (2000) derive the access prices that would result if the firms compete Bertrand, derive the Ramsey-Boiteux prices for this demand and cost structure and show that the access prices that result from Bertrand competition are two high. Indeed, the socially optimal access/termination charge lies below the marginal cost of termination while the (imperfectly) competitive access price lies above the marginal cost of termination. When fixed costs are added to the model, the relationship between the competitive prices and the second-best optimal prices is ambiguous. The results can be further complicated by introducing asymmetries between the competing firms. Various extensions of these models of non-cooperative and cooperative access pricing have recently appeared in the literature. While one might make a case for regulation of access prices in this context, computing the optimal access prices in a two-way access situation would be extremely information intensive and subject to considerable potential for error.

11. Conclusions

For over 100 years economists and policymakers have refined alternative definitions of natural monopoly, developed a variety of different regulatory mechanisms and procedures to mitigate the feared adverse economic consequences of natural monopoly absent regulation, and studied the effects of price and entry regulation in practice. The pendulum of policy toward real and imagined natural monopoly problems has swung from limited regulation, to a dramatic expansion of regulation, to a gradual return to a more limited scope for price and entry regulation. Natural monopoly considerations became a rationale for extending price and entry regulation to industries that clearly did not have natural monopoly characteristics while technological and other economic changes have erased or reduced the significance of natural monopoly characteristics that may once have been a legitimate concern. However, the adverse effects of economic regulation in practice led scholars and policymakers to question whether the costs of imperfect regulation were greater than the costs of imperfect markets. These developments in turn have led to the deregulation of many industries previously subject to price and entry

regulation, to a reduction in the scope of price and entry regulation in several other industries, and to the application of better performance-based regulatory mechanisms to the remaining core natural monopoly segments of these industries.

After the most recent two decades of deregulation, restructuring, and regulatory reform, research on the regulation of the remaining natural monopoly sectors has three primary foci. First, to develop, apply and measure the effects of incentive regulation mechanisms that recognize that regulators have imperfect and asymmetric information about the firms that they regulate and utilize the information regulators can obtain in effective ways. Second, to develop and apply access and pricing rules for regulated monopoly networks that are required to support the efficient expansion of competition in previously regulated segments for which the regulated networks continue to be an essential platform to support this competition. Third, to gain a better understanding of the effects of regulation on dynamic efficiency, in terms of the effects of regulation on the development and diffusion of new services and new supply technologies. These targets of opportunity are being addressed in the scholarly literature but have been especially slow to permeate U.S. regulatory institutions. Successfully bringing this new learning to the regulatory policy arena is a continuing challenge.

References

Ai, C., Martinez, S., Sappington, D.E. (2004). "Incentive regulation and telecommunications service quality". Journal of Regulatory Economics 26 (3), 263–285.

Ai, C., Sappington, D. (2002). "The impact of state incentive regulation on the U.S. telecommunications industry". Journal of Regulatory Economics 22 (2), 133–160.

Armstrong, M., Rochet, J.-C. (1999). "Multi-dimensional screening: a users guide". European Economic Review 43, 959–979.

Armstrong, M., Sappington, D. (2003a). "Recent developments in the theory of regulation". In: Armstrong, M., Porter, R. (Eds.), Handbook of Industrial Organization, vol. III. Elsevier Science Publishers, Amsterdam, in press.

Armstrong, M., Sappington, D.M. (2003b). "Toward a synthesis of models of regulatory policy design with limited information". Mimeo.

Armstrong, M., Sappington, D. (2006). "Regulation, competition and liberalization". Journal of Economic Literature 44, 325–366.

Armstrong, M., Vickers, J. (1991). "Welfare effects of price discrimination by a regulated monopolist". Rand Journal of Economics 22 (4), 571–580.

Armstrong, M., Vickers, J. (2000). "Multiproduct price regulation under asymmetric information". Journal of Industrial Economics 48, 137–160.

Armstrong, M., Cowan, S., Vickers, J. (1994). Regulatory Reform: Economic Analysis and British Experience. MIT Press, Cambridge, MA.

Armstrong, M., Doyle, C., Vickers, J. (1996). "The access pricing problem: a synthesis". Journal of Industrial Economics 44 (2), 131–150.

Averch, H., Johnson, L.L. (1962). "Behavior of the firm under regulatory constraint". American Economic Review 52, 1059–1069.

Bacon, J.W., Besant-Jones, J. (2000). "Global electric power reform, privatization and liberalization of the electric power sector in developing countries". World Bank, Energy and Mining Sector Board Discussion Paper Series, Working Paper No. 2, June.

Bailey, E.E. (1973). Economic Theory of Regulatory Constraint. Heath and Company, Lexington Books, Lexington, D.C.

Bailey, E.E., Coleman, R.D. (1971). "The effect of lagged regulation in an Averch-Johnson model". Bell Journal of Economics 2, 278–292.

Bain, J.S. (1956). Barriers to New Competition. Harvard University Press, Cambridge, MA.

Banerjee, A. (2003). "Does incentive regulation cause degradation of telephone service quality?" Information Economics and Policy 15, 243–269.

Baron, D., Besanko, D. (1984). "Regulation, asymmetric information and auditing". Rand Journal of Economics 15 (4), 447–470.

Baron, D., Besanko, D. (1987a). "Commitment and fairness in a dynamic regulatory relationship". Review of Economic Studies 54 (3), 413–436.

Baron, D., Besanko, D. (1987b). "Monitoring, moral hazard, asymmetric information and risk sharing in procurement contracting". Rand Journal of Economics 18 (4), 509–532.

Baron, D., Myerson, R. (1982). "Regulating a monopolist with unknown costs". Econometrica 50 (4), 911–930.

Baumol, W., Bailey, E., Willig, R. (1977). "Weak invisible hand theorems on the sustainability of prices in multiproduct monopoly". American Economic Review 67 (3), 350–365.

Baumol, W., Klevorick, A.K. (1970). "Input choices and rate of return regulation: an overview of the discussion". Bell Journal of Economics and Management Science 1 (2), 169–190.

Baumol, W., Ordover, J., Willig, R. (1997). "Parity pricing and its critics: a necessary condition for the provision of bottleneck services to competitors". Yale Journal on Regulation 14 (1), 145–164.

Baumol, W., Sidak, G. (1994). "The pricing of inputs sold to competitors". Yale Journal on Regulation 11 (1), 171–202.

Baumol, W.J., Bradford, D.F. (1970). "Optimal departures from marginal cost pricing". American Economic Review 60, 265–283.

Baumol, W.J., Panzar, J., Willig, R.D. (1982). Contestible Markets and the Theory of Industry Structure. Harcourt Brace Javanovich, New York.

Beesley, M., Littlechild, S. (1989). "The regulation of privatized monopolies in the United Kingdom". Rand Journal of Economics 20 (3), 454–472.

Bernstein, J.I., Sappington, D.M. (1999). "Setting the X-factor in price cap regulation plans". Journal of Regulatory Economics 16, 5–25.

Berg, S.V., Tschirhart, J. (1988). Natural Monopoly Regulation: Principles and Practice. Cambridge University Press, Cambridge.

Boiteux, M. (1960). "Peak load pricing". Journal of Business 33, 157–179. Translated from the original in French published in 1951.

Boiteux, M. (1971). "On the management of public monopolies subject to budget constraint". Journal of Economic Theory 3, 219–240. Translated from the original in French and published in Econometrica in 1956.

Bonbright, J.C. (1961). Principles of Public Utility Rates. Columbia University Press, New York.

Borenstein, S. (2005). "Time-varying retail electricity prices: theory and practice". In: Griffin, Puller (Eds.), Electricity Deregulation: Choices and Challenges. University of Chicago Press, Chicago.

Braeutigam, R. (1989). "Optimal prices for natural monopolies". In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. II. Elsevier Science Publishers, Amsterdam.

Braeutigam, R., Magura, M., Panzar, J. (1997). "The effects of incentive regulation on local telephone service rates". Northwestern University, mimeo.

Brennan, T. (1989). "Regulating by capping prices". Journal of Regulatory Economics 1 (2), 133–147.

Brown, S.J., Sibley, D.S. (1986). The Theory of Public Utility Pricing. Cambridge University Press, Cambridge.

Cabral, L., Riordan, M. (1989). "Incentives for cost reduction under price cap regulation". Journal of Regulatory Economics 1 (2), 93–102.

Carlton, D. (1977). "Peak load pricing with stochastic demand". American Economic Review 67, 1006–1010.

Carlton, D., Perloff, J. (2004). Modern Industrial Organization, 4th edn. Addison-Wesley, Boston, MA.
Carrington, R., Coelli, T., Groom, E. (2002). "International benchmarking for monopoly price regulation: the case of Australian gas distribution". Journal of Regulatory Economics 21, 191–216.
Christiansen, L.R., Greene, W.H. (1976). "Economies of scale in U.S. electric power generation". Journal of Political Economy 84, 655–676.
Clark, J.M. (1911). "Rates for public utilities". American Economic Review 1 (3), 473–487.
Clark, J.M. (1913). "Frontiers of regulation and what lies beyond". American Economic Review 3 (1), 114–125.
Clemens, E.W. (1950). Economics of Public Utilities. Appleton-Century-Crofts, New York.
Cowing, T.G. (1974). "Technical change and scale economies in an engineering production function: the case of steam electric power". Journal of Industrial Economics 23, 135–152.
Crandall, R.W., Hausman, J.A. (2000). "Competition in U.S. telecommunications services: effects of the 1996 legislation". In: Peltzman, S., Winston, C. (Eds.), Deregulation of Network Industries. Brookings Institution Press, Washington, D.C.
Crandall, R.W., Waverman, L. (1995). Talk is Cheap: The Promise of Regulatory Reform in North American. Brookings, Washington, D.C.
Crawford, G. (2000). "The impact of the 1992 cable act on consumer demand and welfare: a discrete-choice, differentiated products approach". Rand Journal of Economics 31, 422–450.
Crew, M.A., Kleinforfer, P.R. (1976). "Peak load pricing with a diverse technology". Bell Journal of Economics 7, 207–231.
Crew, M.A., Kleinforfer, P.R. (1986). The Economics of Public Utility Regulation. MIT Press, Cambridge, MA.
Dana, J. (1993). "The organization and scope of agents: regulating multiproduct industries". Journal of Economic Theory 59 (2), 288–310.
Demsetz, H. (1968). "Why regulate utilities". Journal of Law and Economics 11 (1), 55–65.
Dreze, J. (1964). "Contributions of French economists to theory and public policy". American Economic Review 54 (4), 2–64.
Ely, R. (1937). Outlines of Economics. MacMillan, New York.
Estache, A., Kouasi, E. (2002). "Sector organization, governance and the inefficiencies of African water utilities". World Bank Policy Research Working Paper No. 2890, September.
Estache, A., Guasch, J.-L., Trujillo, L. (2003). "Price caps, efficiency payoffs, and infrastructure contract renegotiation in Latin America". Mimeo.
Estache, A., Rossi, M.A., Ruzzier, C.A. (2004). "The case for international coordination of electricity regulation: evidence from the measurement of efficiency in South America". Journal of Regulatory Economics 25 (3), 271–295.
Evans, D.S. (1983). Breaking Up Bell: Essays in Industrial Organization and Regulations. North-Holland, New York.
Farrer, T.H. (1902). The State in Relation to Trade. Macmillan, London.
Faulhaber, G.R. (1975). "Cross-subsidization: pricing in public utility enterprises". American Economic Review 65, 966–977.
Fiorina, M. (1982). "Legislative choice of regulatory forums: legal process or administrative process". Public Choice 39, 33–36.
Fraquelli, G., Picenza, M., Vannoni, D. (2004). "Scope and scale economies from multi-utilities: evidence from gas, water and electricity combinations". Applied Economcs 36 (18), 2045–2057.
Gagnepain, P., Ivaldi, M. (2002). "Incentive regulatory policies: the case of public transit in France". Rand Journal of Economics 33, 605–629.
Gasmi, F., Laffont, J.J., Sharkey, W.W. (2002). "The natural monopoly test reconsidered: an engineering process-based approach to empirical analysis in telecommunications". International Journal of Industrial Organization 20 (4), 435–459.
Giannakis, D., Jamasb, T., Pollitt, M. (2004). "Benchmarking and incentive regulation of quality of service: an application to the U.K. distribution utilities". Cambridge Working Papers in Economics CWEP 0408, Department of Applied Economics, University of Cambridge.

Gilbert, R., Newbery, D. (1994). "The dynamic efficiency of regulatory constitutions". Rand Journal of Economics 26 (2), 243–256.

Gilligan, T.W., Marshall, W.M., Weingast, B.R. (1989). "Regulation and the theory of legislative choice: the interstate commerce act of 1887". Journal of Law and Economics 32, 35–61.

Gilligan, T.W., Marshall, W.J., Weingast, B.R. (1990). "The economic incidence of the interstate commerce act of 1887: a theoretical and empirical analysis of the short-haul pricing constraint". Rand Journal of Economics 21, 189–210.

Glaeser, M.G. (1927). Outlines of Public Utility Economics. MacMillan, New York.

Goldberg, V.C. (1976). "Regulation and administered contracts". Bell Journal of Economics 7, 426–448.

Goolsbee, A., Petrin, A. (2004). "The consumer gains from direct broadcast satellites and the competition with cable television". Econometrica 72 (2), 351–381.

Greene, W.H., Smiley, R.H. (1984). "The effectiveness of utility regulation in a period of changing economic conditions". In: Marchand, M., Pestieau, P., Tulkens, H. (Eds.), The Performance of Public Enterprise: Concepts and Measurement. Elsevier, Amsterdam.

Greenstein, S., McMaster, S., Spiller, P. (1995). "The effect of incentive regulation on infrastructure modernization: local exchange companies' deployment of digital technology". Journal of Economics & Management Strategy 4, 187–236.

Hadlock, C.J., Lee, D.S., Parrino, R. (2002). "Chief executive officer careers in regulated environments: evidence from electric and gas utilities". Journal of Law and Economics 45, 535–564.

Hammond, C.J., Johnes, G., Robinson, T. (2002). "Technical efficiency under alternative regulatory regimes". Journal of Regulatory Economics 22 (3), 251–270.

Hausman, J.A. (1997). "Valuing the effects of regulation on new services in telecommunications". Brookings Papers on Economics Activity: Microeconomics 1–54.

Hausman, J.A. (1998). "Taxation by telecommunications regulation". NBER/Tax Policy and the Economy 12 (1), 29–49.

Hausman, J.A. (1999). "The effects of sunk costs in telecommunications regulation". In: Alleman, J., Noam, E. (Eds.), Real Options: The New Investment Theory and its Applications for Telecommunications Economics. Kluwer Academic, Norwell, MA.

Hausman, J.A. (2002). "Mobile telephone". In: Cave, M.E. et al. (Eds.), Handbook of Telecommunications Economics. Elsevier Science.

Hausman J.A. (2003). "Regulated costs and prices of telecommunications". In: Madden, G. (Ed.), Emerging Telecommunications Networks. Edward Elgar Publishing.

Hausman, J.A., Myers, S.C. (2002). "Regulating the United States railroads: the effects of sunk costs and asymmetric risk". Journal of Regulatory Economics 22, 287–310.

Hausman, J.A., Tardiff, T., Belinfante, A. (1993). "The effects of the breakup of AT&T on telephone penetration in the United States". American Economic Review 83, 178–184.

Hendricks, W. (1977). "Regulation and labor earnings". Bell Journal of Economics 8, 483–496.

Hubbard, T. (2001). "Contractual form and market thickness in trucking". Rand Journal of Economics 32 (2), 369–386.

Hubard, T. (2003). "Information, decisions and productivity: on board computers and capacity utilization in trucking". American Economic Review 94 (4), 1328–1353.

Hughes, T.P. (1983). Networks of Power: Electrification in Western Society 1880–1930. Johns Hopkins University Press, Baltimore, MD.

Isaac, R.M. (1991). "Price cap regulation: a case study of some pitfalls of implementation". Journal of Regulatory Economics 3 (2), 193–210.

Jamasb, T., Pollitt, M. (2001). "Benchmarking and regulation: international electricity experience". Utilities Policy 9, 107–130.

Jamasb, T., Pollitt, M. (2003). "International benchmarking and regulation: an application to European electricity distribution utilities". Energy Policy 31, 1609–1622.

Jarrell, G.A. (1978). "The demand for state regulation of the electric utility industry". Journal of Law and Economics 21, 269–295.

Joskow, P.L. (1972). "The determination of the allowed rate of return in a formal regulatory hearing". Bell Journal of Economics and Management Science 3, 633–644.
Joskow, P.L. (1973). "Pricing decisions of regulated firms". Bell Journal of Economics and Management Science 4, 118–140.
Joskow, P.L. (1974). "Inflation and environmental concern: structural change in the process of public utility price regulation". Journal of Law and Economics 17, 291–327.
Joskow, P.L. (1976). "Contributions to the theory of marginal cost pricing". Bell Journal of Economics 7 (1), 197–206.
Joskow, P.L. (1989). "Regulatory failure, regulatory reform and structural change in the electric power industry". Brookings Papers on Economic Activity: Microeconomic 125–199.
Joskow, P.L. (1997). "Restructuring, competition and regulatory reform in the U.S. electricity sector". Journal of Economic Perspectives 11 (3), 119–138.
Joskow, P.L. (2000). "Deregulation and regulatory reform in the U.S. electric power industry". In: Peltzman, S., Winston, C. (Eds.), Deregulation of Network Industries. Brookings Institution Press, Washington, D.C.
Joskow, P.L. (2005a). "Transmission policy in the United States". Utilities Policy 13, 95–115.
Joskow, P.L. (2005b). "Regulation and deregulation after 25 years". International Review of Industrial Organization 26, 169–193.
Joskow, P.L. (2006). "Incentive regulation in theory and practice". NBER Regulation Project, mimeo. (http://econ-www.mit.edu/faculty/download_pdf.php?id=1220.)
Joskow, P.L., Noll, R.G. (1981). "Regulation in theory and practice: an overview". In: From, G. (Ed.), Studies in Public Regulation. MIT Press, Cambridge, MA.
Joskow, P.L., Noll, R.G. (1999). "The Bell doctrine: applications in telecommunications, electricity and other network industries". Stanford Law Review 51 (5), 1249–1315.
Joskow, P.L., Rose, N.L. (1985). "The effects of technological change, experience and environmental regulation on the costs of coal-burning power plants". Rand Journal of Economics 16 (1), 1–27.
Joskow, P.L., Rose, N.L. (1989). "The effects of economic regulation". In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. II. North-Holland, Amsterdam.
Joskow, P.L., Rose, N.L., Wolfram, C.D. (1996). "Political constraints on executive compensation: evidence from the electric utility industry". Rand Journal of Economics 27, 165–182.
Joskow, P.L., Schmalensee, R. (1983). Markets for Power. MIT Press, Cambridge, MA.
Joskow, P.L., Schmalensee, R. (1986). "Incentive regulation for electric utilities". Yale Journal on Regulation 4, 1–49.
Joskow, P.L., Tirole, J. (2005). "Retail electricity competition". Rand Journal of Economics (in press). (http://econ-www.mit.edu/faculty/download_pdf.php?id=918.)
Joskow, P.L., Tirole, J. (2006). "Reliability and competitive electricity markets". Rand Journal of Economics (in press). (http://econ-www.mit.edu/faculty/download_pdf.php?id=917.)
Kahn, A.E. (1970). The Economics of Regulation: Principles and Institutions, volume I. Wiley, New York.
Katz, M., Shapiro, C. (1986). "Technology adoption in the presence of network externalities". Journal of Political Economy 94, 822–841.
Kaysen, C., Turner, D. (1959). Antitrust Policy: An Economic ad Legal Analysis. Harvard University Press, Cambridge, MA.
Klemperer, P. (2002). "What really matters in auction design". Journal of Economic Perspectives 16, 169–189.
Klevorick, A.K. (1971). "The optimal fair rate of return". Bell Journal of Economics 2, 122–153.
Klevorick, A.K. (1973). "The behavior of the firm subject to stochastic regulatory review". Bell Journal of Economics 4, 57–88.
Kolbe, L., Tye, W. (1991). "The Duquesne opinion: how much 'Hope' is there for investors in regulated firms?" Yale Journal on Regulation 8 (1), 113–157.
Kolko, G. (1965). Railroads and Regulation 1877–1916. Princeton University Press, Princeton.
Kridel, D., Sappington, D., Weisman, D. (1996). "The effects of incentive regulation in the telecommunications industries: a survey". Journal of Regulatory Economics 18, 269–306.

Kwoka, J. (1993). "Implementing price caps in telecommunications". Journal of Policy Analysis and Management 12 (4), 722–756.
Laffont, J.-J. (1999). "Competition, information and development". In: Annual World Bank Conference on Development Economics 1998. The World Bank, Washington, D.C.
Laffont, J.-J., Rey, P., Tirole, J. (1998a). "Network competition: I. Overview and nondiscriminatory pricing". Rand Journal of Economics 29, 1–37.
Laffont, J.-J., Rey, P., Tirole, J. (1998b). "Network competition: II. Price discrimination". Rand Journal of Economics 29, 38–56.
Laffont, J.-J., Tirole, J. (1986). "Using cost observations to regulate firms". Journal of Political Economy 94 (3), 614–641.
Laffont, J.-J., Tirole, J. (1988a). "Auctioning incentive contracts". Journal of Political Economy 95 (5), 921–937.
Laffont, J.-J., Tirole, J. (1988b). "The dynamics of incentive contracts". Econometrica 56 (5), 1153–1176.
Laffont, J.-J., Tirole, J. (1990a). "Adverse selection and renegotiation in procurement". Review of Economic Studies 57 (4), 597–626.
Laffont, J.-J., Tirole, J. (1990b). "Optimal bypass and cream-skimming". American Economic Review 80 (4), 1041–1051.
Laffont, J.-J., Tirole, J. (1993). A Theory of Incentives in Regulation and Propcurement. MIT Press, Cambridge, MA.
Laffont, J.-J., Tirole, J. (1996). "Creating competition through interconnection: theory and practice". Journal of Regulatory Economics 10 (3), 227–256.
Laffont, J.-J., Tirole, J. (2000). Competition in Telecommunication. MIT Press, Cambridge, MA.
Levy, B., Spiller, P. (1994). "The institutional foundations of regulatory commitment: a comparative analysis of telecommunications". Journal of Law, Economics and Organization 10 (2), 201–246.
Lewis, T., Sappington, D.M. (1988a). "Regulating a monopolist with unknown demand". American Economic Review 78 (5), 986–998.
Lewis, T., Sappington, D.M. (1988b). "Regulating a monopolist with unknown demand and cost functions". Rand Journal of Economics 18 (3), 438–457.
Lewis, T., Sappington, D. (1989). "Regulatory options and price cap regulation". Rand Journal of Economics 20 (3), 405–416.
Loeb, M., Magat, W. (1979). "A decentralized method for utility regulation". Journal of Law and Economics 22 (2), 399–404.
Lowry, E.D. (1973). "Justification for regulation: the case for natural monopoly". Public Utilities Fortnightly November 8, 1–7.
Lyon, T. (1996). "A model of the sliding scale". Journal of Regulatory Economics 9 (3), 227–247.
Marshall, A. (1890). Principles of Economics, 8ht edn. MacMillan, London. (1966).
Mathios, A.D., Rogers, R.P. (1989). "The impact of alternative forms of state regulation of AT&T direct-dial, long-distance telephone rates". Rand Journal of Economics 20, 437–453.
McCubbins, M.D. (1985). "The legislative design of regulatory structure". American Journal of Political Science 29, 721–748.
McCubbins, M.D., Noll, R.G., Weingast, B.R. (1987). "Administrative procedures as instruments of corporate control". Journal of Law, Economics and Organization 3, 243–277.
McDonald, F. (1962). Insull. University of Chicago Press, Chicago.
Megginson, W., Netter, J. (2001). "From state to market: a survey of empirical studies of privatization". Journal of Economic Literature 39, 321–389.
Mullin, W.P. (2000). "Railroad revisionists revisited: stock market evidence from the progressive era". Journal of Regulatory Economics 17 (1), 25–47.
Myers, S.C. (1972a). "The application of finance theory to public utility rate cases". Bell Journal of Economics and Management Science 3 (1), 58–97.
Myers, S.C. (1972b). "On the use of β in regulatory proceedings". Bell Journal of Economics and Management Science 3 (2), 622–627.

National Civic Federation (1907). Municipal and Private Operation of Public Utilities, volume I. National Civic Federation, New York.

Nelson, J.R. (1964). Marginal Cost Pricing in Practice. Prentice-Hall, Englewood-Cliffs, N.J.

Newbery, D.M., Pollitt, M.G. (1997). "The restructuring and privatisation of Britain's CEGB: was it worth it?" Journal of Industrial Economics 45 (3), 269–303.

Noll, R.G. (1989). "Economic perspectives on the politics of regulation". In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. II. North-Holland, Amsterdam.

Office of Gas and Electricity Markets (OFGEM) (2004a). "Electricity distribution price control review: policy document". March, London, UK.

Office of Gas and Electricity Markets (OFGEM) (2004b). "Electricity distribution price control review: final proposals". 265/04, November, London.

Office of Gas and Electricity Markets (OFGEM) (2004c). "NGC system operator incentive scheme from April 2005: initial proposals". December, London.

Office of Gas and Electricity Markets (OFGEM) (2004d). "Electricity transmission network reliability incentive scheme: final proposals". December, London.

Owen, B., Brauetigam, R. (1978). The Regulation Game: Strategic Use of the Administrative Process. Ballinger Publishing Company, Cambridge, MA.

Ordover, J.A., Saloner, G., Salop, S.C. (1990). "Equilibrium vertical foreclosure". American Economic Review 80, 127–142.

Palmer, K. (1992). "A test for cross subsidies in local telephone rates: do business customers subsidize residential customers?" Rand Journal of Economics 23, 415–431.

Panzar, J.C. (1976). "A neoclassical approach to peak load pricing". Bell Journal of Economics 7, 521–530.

Peltzman, S. (1989). "The economic theory of regulation after a decade of deregulation". Brookings Papers on Economic Activity: Microeconomics 1–60.

Phillips, C.F. Jr. (1993). The Regulation of Public Utilities: Theory and Practice. Public Utilities Report, Inc., Arlington, VA.

Pindyck, R. (2004). "Pricing capital under mandatory unbundling and facilities sharing". December, mimeo.

Pindyck, R., Rubinfeld, D. (2001). Microeconomics, 5th edn. Prentice-Hall, Upper Saddle River, N.J.

Posner, R.A. (1969). "Natural monopoly and regulation". Stanford Law Review 21, 548–643.

Posner, R.A. (1971). "Taxation by regulation". Bell Journal of Economics and Management Science 2, 22–50.

Posner, R.A. (1974). "Theories of economic regulation". Bell Journal of Economics 5, 335–358.

Posner, R.A. (1975). "The social cost of monopoly and regulation". Journal of Political Economy 83, 807–827.

Prager, R.A. (1989a). "Using stock price data to measure the effects of regulation: the interstate commerce act and the railroad industry". Rand Journal of Economics 20, 280–290.

Prager, R.A. (1989b). "Franchise bidding for natural monopoly". Journal of Regulatory Economics 1 (2), 115–132.

Prager, R.A. (1990). "Firm behavior in franchise monopoly markets". Rand Journal of Economics 12, 211–225.

Ramsey, F. (1927). "A contribution to the theory of taxation". Economic Journal 37, 47–61.

Riordan, M. (1984). "On delegating price authority to a regulated firm". Rand Journal of Economics 15 (1), 108–115.

Rochet, J.C., Tirole, J. (2003). "Platform competition in two-sided markets". Journal of the European Economic Association 1 (4), 990–1029.

Rochet, J.C., Tirole, J. (2004). "Two-sided markets: an overview". Institute d'Economie Industrielle, March, mimeo.

Rose, N.L. (1987). "Labor rent sharing and regulation: evidence from the trucking industry". Journal of Political Economy 95, 1146–1178. December.

Rose, N.L., Joskow, P.L. (1990). "The diffusion of new technology: evidence from the electric utility industry". Rand Journal of Economics 21 (3), 354–373.

Rose, N., Markiewicz, K., Wolfram, C. (2004). "Does competition reduce costs? Reviewing the impact of regulatory restructuring on U.S. electric generation efficiency". MIT CEEPR Working Paper 04-018. (http://web.mit.edu/ceepr/www/2004-018.pdf.)

Rudnick, H., Zolezzi, J. (2001). "Electric sector deregulation and restructuring in Latin America: lessons to be learnt and possible ways forward". IEEE Proceedings Generation, Transmission and Distribution 148, 180–184.

Salinger, M.E. (1984). "Tobin's q, unionization, and the concentration-profits relationship". Rand Journal of Economics 15, 159–170.

Salinger, M.E. (1998). "Regulating prices to equal forward-looking costs: cost-based prices or price-based cost". Journal of Regulatory Economics 14, 149–163.

Sappington, D.M. (1980). "Strategic firm behavior under a dynamic regulatory adjustment process". Bell Journal of Economics 11 (1), 360–372.

Sappington, D.M. (2003). "The effects of incentive regulation on retail telephone service quality in the United States". Review of Network Economics 2 (3), 355–375.

Sappington, D., Ai, C. (2005). "Reviewing the impact of incentive regulation on U.S. telephone service quality". Utilities Policy 13 (3), 201–210.

Sappington, D., Sibley, D. (1988). "Regulating without cost information: the incremental surplus subsidy scheme". International Economic Review 31 (2), 297–306.

Sappington, D., Sibley, D. (1990). "Regulating without cost information: further observations". International Economic Review 31 (4), 1027–1029.

Sappington, D., et al. (2001). "The state of performance based regulation in the U.S. electric utility industry". Electricity Journal, 71–79.

Schleifer, A. (1985). "A theory of yardstick competition". Rand Journal of Economics 16 (3), 319–327.

Schmalensee, R. (1979). The Control of Natural Monopolies. Lexington Books, Lexington, MA.

Schmalensee, R. (1981). "Output and welfare implications of monopolistic third-degree price discrimination". American Economic Review 71, 242–247.

Schmalensee, R. (1989a). "An expository note on depreciation and profitability under rate of return regulation". Journal of Regulatory Economics 1 (3), 293–298.

Schmalensee, R. (1989b). "Good regulatory regimes". Rand Journal of Economics 20 (3), 417–436.

Sharfman, I.L. (1928). "Valuation of public utilities: discussion". American Economic Review 18 (1), 206–216.

Sharkey, W.W. (1982). The Theory of Natural Monopoly. Cambridge University Press, Cambridge.

Sheshinski, E. (1971). "Welfare aspects of regulatory constraint". American Economic Review 61, 175–178.

Sibley, D. (1989). "Asymmetric information, incentives and price cap regulation". Rand Journal of Economics 20 (3), 392–404.

Sidak, G., Spulber, D. (1997). Deregulatory Takings and the Regulatory Contract. Cambridge University Press, Cambridge.

Spence, M. (1975). "Monopoly, quality and regulation". Bell Journal of Economics 6 (2), 417–429.

Spiegel, Y., Spulber, D. (1994). "The capital structure of regulated firms". Rand Journal of Economics 25 (3), 424–440.

Spiller, P. (1990). "Politicians, interest groups and regulators: a multiple principal agent theory of regulation". Journal of Law and Economics 33 (1), 65–101.

Steiner, P. (1957). "Peak loads and efficient pricing". Quarterly Journal of Economics 71 (4), 585–610.

Stigler, G.J. (1971). "The theory of economic regulation". Bell Journal of Economics and Management Science 2, 3–21.

Stigler, G.J., Friedland, C. (1962). "What can regulators regulate: the case of electricity". Journal of Law and Economics 5, 1–16.

Sutton, J. (1991). Sunk Costs and Market Structure. MIT Press, Cambridge, MA.

Tardiff, T., Taylor, W. (1993). Telephone Company Performance Under Alternative Forms of Regulation in the U.S. National Economic Research Associates.

Teeples, R., Glyer, D. (1987). "Cost of water delivery systems: specific and ownership effects". Review of Economics and Statistics 69, 399–408.

Tirole, J. (1988). The Theory of Industrial Organization. MIT Press, Cambridge, MA.
Troxel, E. (1947). Economics of Public Utilities. Rineheart & Company, New York.
Turvey, R. (1968a). "Peak load pricing". Journal of Political Economy 76, 101–113.
Turvey, R. (1968b). Optimal Pricing and Investment in Electricity Supply: An Essay in Applied Welfare Economics. MIT Press, Cambridge, MA.
Vickers, J. (1995). "Competition and regulation in vertically related markets". Review of Economic Studies 62 (1), 1–17.
Vickers, J., Yarrow, G. (1991). "Economic perspectives on privatization". Journal of Economic Perspectives 5, 111–132.
Vogelsang, I. (2003). "Price regulation of access to telecommunications networks". Journal of Economic Literature 41, 830–862.
Vogelsang, I., Finsinger, J. (1979). "A regulatory adjustment process for optimal pricing of multiproduct firms". Bell Journal of Economics 10 (1), 151–171.
Weiman, D.F., Levin, R.C. (1994). "Preying for monopoly? The case of southern Bell Telephone, 1894–1912". Journal of Political Economy 102 (1), 103–126.
Weingast, B.R., Moran, M.J. (1983). "Bureaucratic discretion or congressional control? Regulatory policy-making at the Federal Trade Commission". Journal of Political Economy 91, 765–780.
Weitzman, M. (1980). "The ratchet principle and performance incentives". Bell Journal of Economics 11 (1), 302–308.
Weitzman, M.A. (1983). "Contestable markets: an uprising in the theory of industry structure: comment". American Economic Review 73 (3), 486–487.
Williamson, O.E. (1976). "Franchise bidding for natural monopolies: in general and with respect to CATV". Bell Journal of Economics 7 (1), 73–104.
Williamson, O.E. (1985). The Economic Institutions of Capital: Firms, Markets and Contracting. Free Press, New York.
Williamson, O.E. (1996). The Mechanisms of Governance. Oxford University Press, New York.
Willig, R. (1978). "Pareto-superior non-linear outlay schedules". Bell Journal of Economics 9 (1), 56–69.
Willig, R. (1979). "The theory of network access pricing". In: Trebing, H. (Ed.), Issues in Public Utility Regulation. Michigan State University Press, East Lansing, MI.
Winston, C. (1993). "Economic deregulation: days of reckoning for microeconomists". Journal of Economic Literature 31 (3), 1263–1289.
Winston, C., Peltzman, S. (2000). Deregulation of Network Industries. Brookings Institution Press, Washington, D.C.
Zupan, M. (1989a). "Cable franchise renewals: do incumbent firms behave opportunistically". Rand Journal of Economics 20 (4), 473–482.
Zupan, M. (1989b). "The efficacy of franchise bidding schemes for CATV: some systematic evidence". Journal of Law and Economics 32 (2), 401–456.