January 24, 2008

Is Protecting Sunk Investments by Consumers a Key Rationale for Natural Monopoly Regulation?

Darryl R Biggar
Why regulate natural monopolies? Conventional economic theory points to the price-marginal cost margin and the ensuing deadweight loss. But this hypothesis does a poor job of explaining the way that regulators behave in practice. In particular, regulators have tended to eschew forms of price discrimination, including Ramsey pricing and peak-load pricing, and have actively pursued stable pricing. When pressed, regulators justify rejection of the conventional theory on the grounds of pursuit of “fairness” or “equity”. Could it be that the conventional economic justification for natural monopoly regulation is flawed? Some economists have emphasised the role of regulation in protecting the sunk investments made by regulated firms. But this hypothesis cannot explain the patterns of regulation we observe in practice, such as why regulators seem to focus more attention on keeping regulated prices down than on keeping prices up. This paper proposes an alternative hypothesis: that natural monopoly regulation exists to protect the sunk investments made by consumers of the regulated firm. This hypothesis can explain many of the practices of regulators which make little or no sense under conventional economic theory, such as the desire to pursue stable prices, the aversion to forms of price discrimination such as Ramsey pricing, and the role of incremental cost as a pricing floor. In my view, this hypothesis goes some way to bridging the remaining gap between economic theory and regulatory practice.

---

1 Darryl Biggar is an economist with the Australian Competition and Consumer Commission (“ACCC”) and the Australian Energy Regulator (“AER”). The views expressed in this paper are his own and not the views of the ACCC or the AER.
1. Introduction

When asked to describe the primary source of the economic harm arising from monopoly, economists conventionally point to the margin between price and marginal cost and the resulting “deadweight loss”. For several decades at least, economists have argued that regulators should focus on ensuring that regulated prices – at least at the margin - approximate marginal cost. Where prices must (for cost-recovery purposes, say) depart from marginal cost, economists conventionally recommend that they should do so in a manner which minimises deadweight loss.

But is the ultimate purpose of natural monopoly regulation really the elimination of deadweight loss? Unfortunately, this hypothesis does a poor job of explaining the patterns of regulation that we observe in practice, and the behaviour of regulators themselves.

In fact, there are many policy prescriptions which are soundly based in conventional economic theory but which are partially or entirely rejected in regulatory practice. This includes for example, the efficiency of perfect price discrimination, the value of Ramsey pricing, and the benefits of peak-load pricing. Yet, in each case, these recommendations of economists are almost always partially or totally rejected by regulators in practice. Is this because economists are not preaching loud enough? Or, is something missing from the conventional economic theory?

I suggest that the conventional economic rationale for natural monopoly regulation is flawed. Rather than minimization of deadweight loss, I suggest that the primary rationale for natural monopoly regulation is the protection of sunk investments – in particular, the sunk investments made by the customers of the regulated firm.

In the case of most monopoly services, users can choose to make sunk investments which increase their demand for or value of the monopoly services – such as choosing where to live, where to locate their manufacturing plant, or whether to invest in developing new products which make use of the monopoly services. The need for sunk investments gives rise a conventional hold-up problem – users fear that once these investments are made, the value of the investment will be expropriated by the monopolist. Although there are some partial mechanisms for protecting and promoting sunk investments by users (which are discussed later), the most straightforward approach is through conventional natural monopoly regulation.

There is a small but significant strand of the economics literature on regulation which, drawing on the transactions-cost school, argues that regulation is best understood as preventing hold-up and thereby promoting investment in sunk assets. However, this literature has focused on promoting sunk investment by the regulated firm. I argue that this approach cannot explain the existence of the regulatory controls in the first place. In contrast, I focus here on promoting sunk investment by the customers of the regulated firm.

I suggest that viewing natural monopoly regulation as being designed to protect user investment allows us to better understand and explain the patterns of regulation and the behaviour of regulators. In particular, it allows us to explain why some of the key policy prescriptions of conventional economic theory have been ignored in practice. Importantly, it allows us to explain persistent patterns of behaviour by regulators without recourse to arguments based on “equity” or “fairness”.

This paper has six parts. The next part explores the traditional economic objective for natural monopoly regulation – the elimination of deadweight loss – and explores whether this objective can explain the patterns of regulation we observe in practice. The third part looks in more detail at whether regulation can be explained as protecting the sunk investment of the regulated firm. The fourth part sets out the alternative rationale put forward here: the protection and promotion of sunk user investment. The fifth part tests this hypothesis by exploring how well it explains the patterns of regulation observed in practice. The sixth part concludes.
2. **Is natural monopoly regulation just about market power and deadweight loss?**

According to a widespread conventional wisdom in economics, the primary harm from natural monopoly relates to the exercise of market power. Specifically, it is argued that a natural monopoly, selling at a simple linear price and facing a downward-sloping demand curve for its product, will choose a price-quantity combination at which the market price is above the marginal cost of the production of the last unit of output. This gap between the price and marginal cost reduces overall social welfare – relative to the theoretically efficient level – by an amount known as the “deadweight loss”.\(^2\) As summarised by Crocker and Masten (1996):

“The conventional wisdom ... was that public utility regulation was a necessary response to natural monopoly cost conditions. According to the natural monopoly rationale, (i) productive efficiency in an industry with declining long-run average costs requires that a single firm service the market, but (ii) a sole supplier will restrict output to the monopoly level, generating a deadweight loss. Efficiency can only be achieved in such settings if government restricts entry and regulates the price the monopolist can charge to the public”.\(^3\)

This conventional wisdom is typically illustrated in a diagram such as the one below:\(^4\):

\[\text{Diagram showing the deadweight loss due to monopoly.} \]

But is natural monopoly regulation really nothing more than the control of market power? Is the primary rationale for regulation the minimization of dead-weight loss?

To explore this question let’s adopt a “positive” perspective. That is, let’s take the observed patterns of regulatory practice as the phenomenon to be explained and let’s ask the question: Can the hypothesis that the primary rationale for regulation is the minimization of dead-weight loss adequately explain the patterns of regulation we observe in practice?\(^3\)

---

\(^2\) Also known as the “welfare triangle” or “Harberger triangle”.

\(^3\) Crocker and Masten (1996), page 10.

\(^4\) A version of this diagram can be found in most standard textbooks. See, for example, figure I-8 in Train (1991), or Figure 2.5 (page 34) in Church and Ware (2000). See also figure 23.4 (page 1300) in Braeutigam (1989).
Why don’t regulators pursue marginal-cost pricing?

The first and simplest observation we can make is the following. If regulation was solely concerned with the elimination or minimization of deadweight loss, regulators would focus primarily – if not exclusively – on the level of marginal prices and their relationship to marginal costs. Regulators would only care about the level of the firm’s overall revenue and overall costs to the extent that it is necessary to ensure financial viability of the regulated firm.

Economists have for many decades argued the benefits of setting public utility tariffs on the basis of marginal cost. This view is expressed in many classic economic texts on regulation. For example:

- William Vickery (1955) writes:
  
  “[M]arginal cost must play a major and even a dominant role in the elaboration of any scheme of rates or prices that seriously pretends to have as a major motive the efficient utilization of available resources and facilities … One may, for various good and sufficient reasons, hesitate to embrace marginal cost pricing in all of its ramifications as an absolute standard. But no approach to utility pricing can be considered truly rational which does not give an important and even a major weight to marginal cost considerations.”

- Alfred Kahn in his classic text, *The Economics of Regulation* (1970), emphasises:
  
  “The central policy prescription of microeconomics is the equation of price and marginal cost. If economic theory is to have any relevance to public utility pricing, that is the point at which the inquiry must begin”.

However, despite the emphasis by economists, the central role of marginal cost in pricing seems to be downplayed in practice. Any focus on the relationship between marginal prices and marginal costs seems to be of secondary importance at best. Bonbright et al (1988) summarises:

“While most people in the public utility community are aware of and would probably acknowledge the validity of marginal cost pricing, many would minimize it in actual ratemaking on grounds of either practicality or of a lack of singlemindedness to economic efficiency …

It is no secret that ratemaking in the United States has historically deviated significantly from the first-best marginal cost ideal.”

---


6 Kahn (1970), volume I, page 65, emphasis added. Precisely the same view arises in more recent texts. For example, Joskow (1991), page 69, writes: “An efficient set of rate-making rules should satisfy two sets of constraints. First, these rules should have the property that the expected present discounted value of future cash flows is equal to the cost of the investments that utilities must make to provide economical and reliable service. Second, at every point in time, prices should reflect the marginal economic cost of providing the relevant services, reflecting changes in cost and demand conditions over time”. Joskow does not explain why the first condition is necessary. Train (1995): “The central issue of regulatory economics is the design of mechanisms that regulators can apply to induce firms to achieve optimal outcomes. … In many situations, this [optimal outcome] is a direct application of concepts from micro-economic theory, such as that price equals marginal cost at the optimal output level”. Viscusi, Vernon and Harrington (1995), page 358-359, “The most obvious candidate for the efficient price is, of course, marginal cost. A natural monopolist that charges marginal cost for each product is said to practice linear (or uniform) marginal cost pricing”.

7 Bonbright (1988), page 415. Bonbright goes on: “This has been due to a variety of factors: (1) Rates were based on average historic costs which were sharply below the current costs; (2) Rates were based on private
In fact, regulators seem reluctant to promote the forms of price discrimination necessary for marginal cost pricing, especially when that might allow the regulated firm to earn revenues in excess of its costs.

For example, let’s take the extreme case of perfect price-discrimination. Under theoretical first-degree or “perfect” price-discrimination, the monopolist charges for each unit of service the total willingness-to-pay of the consumer. Perfect price discrimination ensures that consumers continue to consume up to the point where their marginal value (or marginal willingness-to-pay) of the last unit consumed is just equal to the marginal cost – eliminating the deadweight loss and maximising allocative efficiency. This is usually illustrated in a diagram such as the following:

In practice, it is rare for a monopolist to have sufficient information on consumer willingness-to-pay to be able to implement a strategy of perfect price discrimination. Nevertheless, if the elimination of deadweight loss were the sole rationale for regulation, regulators should welcome perfect price discrimination, or attempts to move towards perfect price discrimination, in those situations where it is feasible, no matter what the level of the monopolist’s earnings. This seems counter to the behaviour we would expect in practice. It seems that most regulators and policy-makers would still have strong concerns about a monopolist that could receive the entire producers’ and consumers’ surplus, even if the pricing were fully “efficient”. According to standard economic theory perfect price discrimination is welfare maximising, yet it seems likely to be uniformly rejected in regulatory practice.

---

In (1991) explains this as follows: “Under primary price discrimination, the firm extracts all surplus: it charges each customer exactly the customer’s willingness to pay. Because the firm obtains all surplus as profit, profit maximization and surplus maximization are identical: the firm naturally chooses the optimal outcome. This transfer of all surplus to the firm is clearly the most straightforward way (at least theoretically) to provide consistency between the regulator’s goal and the firm’s profit motive”. Train (1991), page 90. Leeson and Sobel (2006) write: “Is perfect price discrimination socially efficient? A survey of economists, or glance through any economics textbook, would universally suggest the answer is ‘yes’. … The perfectly price discriminating monopolist resolves the deadweight loss by expanding output, creating a social gain, and transfers consumer surplus to himself, which is socially neutral. The net result is increased social welfare”.

---

8 Train (1991) explains this as follows: “Under primary price discrimination, the firm extracts all surplus: it charges each customer exactly the customer’s willingness to pay. Because the firm obtains all surplus as profit, profit maximization and surplus maximization are identical: the firm naturally chooses the optimal outcome. This transfer of all surplus to the firm is clearly the most straightforward way (at least theoretically) to provide consistency between the regulator’s goal and the firm’s profit motive”. Train (1991), page 90. Leeson and Sobel (2006) write: “Is perfect price discrimination socially efficient? A survey of economists, or glance through any economics textbook, would universally suggest the answer is ‘yes’. … The perfectly price discriminating monopolist resolves the deadweight loss by expanding output, creating a social gain, and transfers consumer surplus to himself, which is socially neutral. The net result is increased social welfare”.
When pushed to explain why this particular policy recommendation is uniformly set aside, economists resort to non-economic or quasi-economic arguments, such as ethical or distribution objections. For example:

“[P]rimary price discrimination results in the firm making large profits, consisting of the entire surplus. Thus, even though the optimal output is attained, the benefits of attaining this output all accrue to the monopolist. The regulator might consider this distribution of benefits to be inequitable …

It might be considered inequitable for different customers to pay different prices for the same goods or services. That is, the basic premise of price discrimination might conflict with the regulator’s goals regarding equity, even though the regulator’s goals regarding input and output are met”.

“The distributive effect of [perfect price] discrimination is very adverse to consumers as a whole: they get no part of the wealth that is created (or only a small piece of it in the case of imperfect discrimination). It may seem very unfair that the producer takes all”.

But these responses are inadequate. On closer inspection they merely raise additional questions: Why, exactly, is it “inequitable” for different customers to pay different prices? Price discrimination may in fact result in wealthier customers paying more for the monopoly service than poorer customers. Couldn’t this be said to be more “equitable”? In fact, price discrimination may result in poorer customers receiving significantly larger quantities of the monopoly service at a cheaper price than if the monopolist was forced to charge a single price to all customers. Is this inequitable? Furthermore, in intermediate-goods industries the primary customers of a regulated firm are often themselves large firms (consider, for example, the above-rail customers of a regulated rail track infrastructure). What exactly are the equity or distribution arguments for preventing discrimination in this case?

Similarly, why exactly is it inequitable for the total surplus to pass to the monopolist? Is it something to do with concerns about the resulting distribution of income? What if the government issued shares in the price-discriminating monopolist to lower-income households? Would this entirely eliminate the rationale for regulation? Intuition suggests that the answer is no. Regulatory objections to perfect price discrimination cannot be based on distaste for the income distribution implications alone.

Overall, as should be clear, the hypothesis that natural monopoly regulation is primarily about the elimination of deadweight loss cannot explain the universal rejection of perfect price discrimination, in practice.

But it is not only perfect price discrimination that is rejected by regulators. For example, consider the use of a two-part tariff, with the “fixed” part varying between customers, and the “variable” part equal to marginal cost. As long as the variable part of the tariff is equal to marginal cost and the fixed part is not so high as to induce some customers from consuming at all, the fully conventionally “efficient” outcome is achieved.

Let’s suppose that a regulated firm approached a natural monopoly regulator with a proposal under which it committed itself to ensuring that its marginal tariffs were always precisely equal to its marginal cost, while ensuring that the fixed tariffs were not so high as to induce some

---

9 Train (1991), page 92-93, emphasis added. Pittman (2007) mounts an argument that these distributional effects are a legitimate concern of economists. His argument is based on the observations that (a) there is a declining marginal utility of income; and (b) despite widespread participation in shares through, say, pension funds, the ownership of firms is disproportionately concentrated on the wealthiest individuals.

10 Leveque (2003), page 21, emphasis added.

11 Consider, for example, the charging practices of the sole doctor operating in a small town.
customers to cease consuming. If this commitment were credible, conventional economic theory suggests the regulator should welcome this proposal and subject to no further regulation at all.

Yet, again, it seems unlikely that any regulator would, in practice, be completely indifferent to the other (non-marginal) components of the tariff of the firm, and indifferent to the size of the profit earned by the firm. It seems likely that any such proposal would, in practice, be rejected. Again, we see that the hypothesis that natural monopoly regulation is primarily about the elimination of deadweight loss does not a poor job of explaining the patterns of regulation we observe in practice.

Furthermore, in some cases the deadweight loss may be small or zero because the demand curve is highly inelastic. Again, in these cases, conventional economic theory would predict that no – or only limited - monopoly regulation would be required. For example, suppose that the monopolist could credibly demonstrate to the regulator that all of its prices were located in a portion of the demand curve which was perfectly inelastic in the region down to marginal cost. If this were the case, the prices would result in precisely the same quantity demanded as if the firm were setting its prices at marginal cost, as in the following diagram:

![Diagram showing marginal cost and demand curves.](image)

According to conventional economic theory, this monopolist should be subject to no further regulation (other than the constraint that it not raise its prices above the point where the quantity demanded starts to reduce). Yet, again, it seems hard to believe that a regulator in practice would leave this firm earning substantial “monopoly rents”, even though there is no loss in economic efficiency as conventionally defined. Again we see that the hypothesis that natural monopoly regulation is primarily focused on the elimination of deadweight loss does a poor job of explaining the behaviour of regulators in practice.

Why don’t regulators pursue Ramsey pricing?

The conventional economic theory of natural monopoly regulation observes that marginal cost pricing may result in the monopoly firm making a loss. If the firm has no other source of funds to cover that loss (such as direct funding from government), the regulator is forced to depart

---

12 Bonbright (1988), page 103, uses this example to argue that there must be an “income redistribution” role for regulation: “In our hypothetical example of residential service for which the demand is completely unresponsive to changes in price within a range from X to 5X per kilowatt-hour (i.e., demand is completely own-price inelastic) the competition version of the compensatory-income transfer function of rates would require that the rate be set at X, since this price suffices to cover cost including a capital-attracting rate of return”.

---
from marginal cost pricing in such a way as to allow the firm to earn sufficient revenue to cover its total costs. However, conventional economic theory still focuses heavily on the deadweight loss. According to conventional economic theory, the regulator should depart from marginal cost pricing in such a way that minimizes the resulting deadweight loss. This implies, of course, Ramsey-Boiteux pricing.

The key policy prescription of Ramsey pricing is that the regulator should take into account the elasticity of demand when departing from marginal cost. The regulated firm should be allowed to raise its prices above marginal cost by a larger margin on those services which are most inelastic, while keeping prices close to marginal cost on those services which are most elastic. By construction, under Ramsey pricing, the regulated firm earns no monopoly rents overall.

Although there may be practical difficulties with implementing Ramsey pricing (such as difficulty obtaining information on the elasticity or shape of the demand curve, as discussed below), in principle, within the conventional economic theory of regulation, Ramsey pricing is a full and complete solution to the pricing question. Ramsey pricing ensures that the deadweight loss is minimised while ensuring that the monopolist earns no excess returns overall.13 If the hypothesis that regulation is primarily about minimizing deadweight loss is correct, Ramsey pricing should be universally acknowledged and accepted by regulators as the correct mechanism for setting prices, at least in principle, if not in actual practice.

In reality, however, Ramsey pricing could, at best, be described as having a lukewarm response by regulators. Despite the large number of regulatory decisions made each year, identifying decisions based explicitly on Ramsey principles is difficult.14 Some regulators have allowed regulated firms flexibility to set their tariffs subject to a weighted average price cap. In theory, under certain conditions, this induces the regulated firm to select Ramsey-like prices – but in practice, the evidence that Ramsey-like prices have emerged is inconclusive.15

In any case, regulators routinely supplement such pricing flexibility with side constraints which limit the rate at which prices can be “rebalanced”. If allowing pricing flexibility induces efficient Ramsey pricing, why would regulators limit the rate at which prices could be rebalanced? Shouldn’t regulators encourage firms to accelerate the rate at which they rebalance, in order to extract the maximum efficiency as quickly as possible?

One commentator sums up the status quo succinctly as follows: Ramsey-Boiteux pricing is “loved by economists but spurned by regulators”.16

As emphasised above, Ramsey pricing is solidly based within conventional economic theory. If that theory correctly captures all the factors that are relevant for regulators, regulators should be flocking to Ramsey pricing. But, as a matter of observation, the lack of enthusiasm for Ramsey pricing in practice is striking.

13 Faulhaber and Baumol (1988) mention “the [economics] profession’s general (but not perfectly complete) acceptance of Ramsey pricing as the theoretically correct rule for regulation of the prices of a multiproduct monopolist”, page 595.

14 For example, see Decker (2007): “… regulatory agencies have typically eschewed Ramsey-based pricing approaches in practice…”; “…despite the general reluctance of regulatory agencies to implement [Ramsey pricing] in practice…”; “Once again, these decisions have unanimously avoided the introduction of Ramsey based pricing structures…” (emphasis added).

15 “However, the evidence on whether Ramsey style price structures have, in fact, developed under price-cap regimes in practice is inconclusive. For example, it has been argued that the price structures that have evolved in different regulated industries in the UK following the wave of privatisation since the 1980’s are less than Ramsey-efficient, and that considerable scope exists to improve overall efficiency by firms rebalancing their tariffs according to Ramsey principles”. Decker (2007), page 10.

16 Albon, Rob, personal communication, 20 August 2007.
Why, exactly? Textbooks suggest various reasons why regulators may be unwilling to pursue Ramsey pricing. The single most common argument given is that it is just too difficult for regulators to obtain the necessary information.\textsuperscript{17} However, as Decker (2007) points out, this claim doesn’t seem to stand up to scrutiny. Regulators routinely deal in areas where they must put effort into gathering and verifying key information. Is it so much harder to assess the magnitude of marginal cost or the elasticity of demand for different services?\textsuperscript{18}

Textbooks also routinely point out that Ramsey pricing may not be adopted due to distributional concerns. It is theoretically possible that the most inelastic services are purchased primarily by low-income customers. It seems somehow inequitable to raise the price-cost margin by the largest amount on the poorest consumers.\textsuperscript{19} But this observation doesn’t explain why Ramsey pricing is not pursued in other cases, where the elasticity of demand is higher on low-income consumers. We do not observe regulators using Ramsey pricing when it seems to accord with these equity concerns and not otherwise. Rather, as already noted, we just don’t observe regulators employing Ramsey pricing at all.\textsuperscript{20}

In any case, when the regulators themselves attempt to articulate why they reject Ramsey pricing, they do not explicitly invoke distributional concerns. Rather they refer to objectives such as “protecting captive customers” or “wider social objectives”.\textsuperscript{21}

Overall, as regards Ramsey pricing, we again find that the hypothesis that natural monopoly regulation is primarily concerned with minimizing deadweight loss is soundly rejected when confronted with the evidence of regulatory practice.

\textit{Why do policy-makers object to price-discrimination?}

Conventional economic theory also has difficulty explaining legislative or regulatory controls on price discrimination more generally. Conventional economic theory is quite clear that some forms

\textsuperscript{17} For example, Faulhaber and Baumol (1988) cite a 1985 decision by the Interstate Commerce Commission, which concluded: “Ramsey pricing is based on a mathematical formula which requires both marginal cost and the elasticity of demand to be quantified for every movement in the carrier’s system. Thus, the amount of data and degree of analysis required seemed overwhelming. We concluded that while Ramsey pricing is useful as a theoretical guideline, it is too difficult and burdensome for universal application”.

\textsuperscript{18} Decker (2007), page 13 writes: “More generally, this ‘too difficult’ explanation in itself raises a number of questions: Are the difficulties associated with the estimation of elasticity estimates in respect of Ramsey prices insurmountable, or are there ways to design estimation models which address the concerns? Is it possible to learn from how elasticity estimates are developed in other areas of antitrust or in the academic community? Is the standard being applied here consistent with other areas of regulatory practice where there is substantial uncertainty, for example cost estimation generally? What is the magnitude of the potential risks associated with adopting broad, but imperfect, elasticity estimates, and how does this risk compare to that associated with alternative and simpler approaches? Is it really the case that elasticity estimates are difficult to estimate in all regulated industries – for example, compare postal services and mobile phone services?”.

\textsuperscript{19} See Church and Ware (2000), page 796.

\textsuperscript{20} Ramsey prices are also sometimes criticised because they are not necessarily above incremental cost and therefore are not necessarily “subsidy-free” (in the sense of Faulhaber, 1975). But this merely raises the question why an incremental cost floor is a relevant economic consideration. This is discussed further later in the paper.

\textsuperscript{21} As cited in Decker (2007), page 16: “Postcomm has a specific duty in relation to vulnerable users at an early stage of competition and believes that it is important to protect them from Royal Mail reacting to competition by reducing the price of competitive products through recouping revenues through increasing prices of captive customers (Postcomm 2006)” and “The notion of Ramsey pricing implies that the mark-up should be higher in those markets where demand is less responsive to changes in price…. it would be necessary to consider carefully whether Ramsey pricing would have any significant implications for Ofgem’s wider social and environmental objectives. (Ofgem 2000)”
of price discrimination are welfare improving. We have already seen two examples above – the (admittedly extreme) case of perfect price discrimination, and the use of a two-part tariff, with a variable charge equal to marginal cost. Price discrimination can be welfare-improving precisely when it reduces the deadweight loss associated with the exercise of market power. Ramsey pricing is one such form of efficient price discrimination.

In practice, however, we systematically find that regulators either choose to or are required to limit the extent of price discrimination. For example, a staple of public utility regulation in the US is the requirement that public utility rates must be just and reasonable, and not unduly discriminatory:

“One of the most nearly universal obligations imposed by federal and state laws on public utilities is the obligation to furnish service and to charge rates that will avoid undue or unjust discrimination among customers, actual or potential”.

In the EU, according to Leveque (2003), there are similar prohibitions on discrimination in the electricity regulations:

“The principle [of non-discrimination] can be ranked as the most important [principle] if one refers to the number of times it is mentioned in the electricity statutes”.

If price discrimination is economically efficient (according to conventional economic theory) why is it rejected or strictly circumscribed in practice? The hypothesis that the primary rationale of regulation is the minimization of deadweight loss doesn’t seem to explain the patterns of market intervention we observe in practice.

Why? Is it that economists are not being heard? Is it that legislators do not have the ability or the willingness to understand economics? Or is it, as I suggest below, that there is a fundamental flaw in the conventional economic approach to market power?

Why is there so much resistance to “peak load pricing”?

The failure of the conventional economic approach to regulation is also apparent in the response by regulators to another favourite of economists: peak-load pricing. Peak-load pricing is a form of marginal-cost pricing applying in those industries in which demand which varies more quickly than the capacity of the network. In this context, standard theory shows that it makes sense to differentiate tariffs by time of use, with tariffs at off-peak time reduced down to marginal cost, while tariffs at peak times are increased to ration demand to match the available network capacity. In effect, all of the fixed and common costs of the network are recovered in the charges at peak times.

---

22 Bonbright (1988), page 515. Bonbright (1988), page 517 goes on: “By and large, the tribunals have shared a popular feeling that ‘discrimination’ as a practice of ratemaking, is an odious and pejorative term. This virulent sentiment has made regulators reluctant to characterize as discriminatory any rates which they find lawful”. Joskow (1991), page 69, seems to argue that rules against discrimination are designed to prevent regulators using their pricing powers to impose what amounts to implicit taxes and subsidies: “Since customers cannot readily turn to competing suppliers of the services provided by the public utility, and resale of public utility services is either very costly or illegal, substantial opportunities exist for price discrimination to achieve political ends. … [T]he public utility rate-making process has always represented a tantalizing target of opportunity for redistributing income by hiding implicit taxes and subsidies in public utility rates. It is an opportunity that regulators have taken advantage of quite frequently”.

23 Leveque (2003), page 15. Laffont and Tirole (1993), footnote 52 observe that the “rule of non-discrimination among consumers … has strangely been interpreted by charging identical prices to consumers with vastly different marginal cost of service (e.g., city and rural customers)”.

24 Kahn (1970) summarises: “The economic principle here is absolutely clear: if the same type of capacity serves all users, capacity costs as such should be levied only on utilization at the peak. Every purchase at that time makes its proportionate contribution in the long-run to the incurrence of those capacity costs and
Various forms of time-of-use pricing are routinely adopted in a number of regulated industries, including, of course, telecommunications and electricity. However, it is not clear that the nature of that time-of-use pricing always follows economic principles and, in any case, in some industries (for example, in road pricing), peak load pricing has met stiff resistance. Faulhaber and Baumol (1988) write:

“Prices differentiated by time of day and day of week have proliferated in telecommunications since the 1960s; however, the methods used to determine peak and off-peak prices owe little to economists’ formulas and more to accountants’ methods in allocating capacity costs. In other areas of public utility application, such as transportation, peak-load pricing has found little application, possibly because of a preoccupation with ‘fairness’ …”.

Kahn (1970) is explicit about the resistance of regulators to wholeheartedly adopt peak-load pricing:

“Although most public utility executives and regulators recognize that peak responsibility pricing has some validity, probably most would also vigorously resist its wholehearted acceptance”.

If regulation is primarily concerned with maximising economic welfare (as conventionally defined), why do regulators resist peak-load pricing in practice? Again we see that the “deadweight loss” hypothesis simply does a poor job of explaining the patterns of regulation we observe in practice. Could it be that regulators are just bad economists? Or is there something wrong with the conventional economic rationale for natural monopoly regulation?

Overall, my conclusion is that either “the minimization/elimination of deadweight loss” or the “maximization of economic welfare” (as economic welfare is usually defined) is inadequate as a positive explanation of the primary rationale for regulation. There must be other concerns in mind by regulators when they make their decisions. The fundamental rationale for regulation in the real world (as opposed to the world of economics textbooks) must lie elsewhere. But what might be this alternative rationale for regulation? This is the question we consider next.
3. Alternative rationales for natural monopoly regulation

If the primary rationale for natural monopoly regulation is not the control of market power (defined as pricing above marginal cost), then what might it be? Other possible rationales include the following:

- To encourage the productive efficiency of the monopolist.
- To eliminate the incentive to waste resources seeking to obtain a position of monopoly.
- To protect the sunk investment of the monopolist.

Let's explore each of these in turn. I will deal with the first two possibilities quickly, before discussing the third rationale in the bulk of this section.

Could it be that the primary rationale for regulation is to promote the efficient provision of the monopoly service? After all, as a matter of observation regulators seem to put a lot of effort into maintaining and strengthening incentives for efficient operation by the regulated firm.

But, why is regulation to maintain productive efficiency necessary? The normal governance mechanisms on privately owned firms go some way towards ensuring that all privately-owned firms – even monopoly firms – have some incentive to minimise their costs, so as to maximise their profits. Concerns have been expressed by some economists that these mechanisms may not work so well in the case of monopoly firms. This has been called “x-inefficiency” or the “quiet life hypothesis.”

This might arise, for example, due to a lack of sufficiently similar comparator firms. In this context, the argument runs, external owners and shareholders cannot compare the monopoly firm with others in order to isolate an effective signal of the performance of the management. As a result, these external owners find it difficult to impose sufficiently strong performance incentives on management, leading to lower overall efficiency.

This argument suggests one possible reason why monopoly firms might be somewhat less efficient than other firms in the economy, but can we say that the promotion of productive efficiency by the regulated firm is the primary rationale for regulation?

If productive efficiency is the objective, is regulation the best mechanism to achieve it? Many commentators argue that natural monopoly regulation hinders rather than promotes incentives for productive efficiency.

More importantly, a monopoly firm may not have any rivals in its own market, but there may be a large number of neighbouring similar monopoly firms. For example, there are hundreds of distribution networks in the US, each with its own local monopoly. Many studies have been carried out comparing the performance of these firms against one another. Presumably the owners of monopoly distribution networks could design quite effective contracts on their managers, rewarding them for improving their productive efficiency relative to their neighbors. Is regulation needed to improve the productive efficiency of these businesses?

The fact that there are regulated firms both with and without good quality comparator firms suggests, that the control of x-inefficiency is not the primary rationale for regulation.\(^\text{28}\)

---

\(^\text{27}\) See, for example, Church and Ware (2000), page 145.

\(^\text{28}\) Consider the following simple thought experiment. Let’s suppose we have two scenarios. In the first scenario we have a monopoly firm (say, an electricity distribution network) with substantial market power but with a large number of virtually identical comparator firms (each with its own monopoly in other regions) and for which the normal commercial governance mechanisms ensure a relatively high degree of productive efficiency. In the second scenario, we have an identical monopoly firm but without any
Could it be that the primary reason for control of a natural monopoly is to prevent resources being wasted in acquiring the monopoly rents? It is true that the opportunity to obtain a degree of market power – and the associated monopoly rents – is a major spur for the expenditure of resources. In some cases this “investment” is duplicative and wasteful. (In other cases, that investment is socially beneficial and is specifically encouraged by public interventions – such as intellectual property rights). But if there were a risk of resources being wasted in seeking to obtain a monopoly position, governments could simply auction a franchise for the monopoly. In this way, all of the monopoly rent would accrue to the government, allowing it to reduce other distortionary taxes while eliminating any incentive to waste resources acquiring monopoly rents. The fact that this solution is not used suggests that eliminating the incentive to waste resources acquiring monopoly rents is not the primary rationale for monopoly regulation.

Natural monopoly regulation and sunk investment by the regulated firm

Could it be that the primary rationale for natural monopoly regulation is to protect the sunk investment of the natural monopoly firm? There is a small but very significant strand of the economic literature which, drawing on the literature on transactions costs, argues precisely this point: that the primary task of the regulator is protecting the sunk investment of the monopolist against the risk of “hold-up” by the regulator. Joskow (1991) explains this perspective as follows:

“To fulfill its obligation to serve, the utility must make substantial investments in long-lived plant and equipment that is highly immobile and has little value in alternative uses. … The combination of franchise-specific sunk investments and franchise exclusivity gives the regulatory agency (or more generally the political process to which it responds) potential power to hold up the utility … Once a public utility has made sunk investments in facilities, it is open to being held up by regulators trying to keep prices as low as possible”.

Jose Gomez-Ibanez and Pablo Spiller have also made major contributions to this literature. Spiller and Tommasi (2005) argue:

“[T]he overarching problem driving the regulation of utilities, whether public or private, and thus the issues politicians have to deal with, is how to limit governmental opportunism, understood as the incentives politicians have to expropriate — once the investments are made — the utilities’ quasi-rents, whether under private or public ownership, so as to garner political support.”

comparator firms for which the normal commercial governance mechanisms may not work effectively. Would we predict that the monopoly would be regulated in the latter scenario but not in the first? That seems unlikely. The conclusion must be that natural monopoly regulation is not primarily about ensuring productive efficiency of the monopoly firm.

Joskow (1991), page 67-68. Crocker and Masten (1996), page 30, similarly write “Foremost among the features making long-term contracts desirable is the opportunistic potential inherent in a small-numbers exchange setting occasioned by relationship-specific investments. Given the need to make such investments, the governance of public utility transactions becomes less a problem of discovering the right price than of identifying the governance structure that economizes on the cost of arriving at and adjusting the terms of trade over the course of that relationship”.


Spiller and Tommasi (2005), page 5, emphasis added. They go on to explain that there are key features of “public utilities” which give rise to this risk of “governmental opportunism”: “The reason for the politicization of infrastructure pricing is threefold. First, the fact that a large component of infrastructure investments is sunk, implies that once the investment is undertaken the operator will be willing to continue operating as long as operating revenue exceed operating costs. Since operating costs do not include a return on sunk investments (but only on the alternative value of those assets), the operating company will be willing to operate even if prices are below total average costs. Second, economies of scale imply that in most utility services, there will be few suppliers in each locality. Thus, the whiff of monopoly will always
This strand of the economics literature is important and has contributed to understanding certain key issues in the design of regulatory institutions and the choice between regulatory governance approaches. Joskow (1991) also emphasises that this line of thinking seems closer to explaining what regulators actually do than the conventional economic focus on deadweight loss:

“The evolution of public utility rate-making and accounting rules bears little if any relationship to the traditional static second-best pricing problem that appears in the academic literature. Instead, the evolution of these accounting and rate-making rules is more closely related to the standard transaction cost economics problem of finding a set of contracting rules that will induce efficient levels of investment, guard against holdups to support these investments, and provide for efficient adaptation to changing economic conditions. The development of twentieth century public-utility accounting and pricing rules was heavily influenced by concerns about encouraging efficient investment, supporting those investments with an adequate but not excessive stream of cash flows and encouraging efficient operation of capital facilities. It was much less concerned with setting prices that matched exactly changing supply and demand conditions at every point in time. On average, prices were supposed to cover total capital and operating costs over time, but prices did not have to equal either marginal or average costs … at each point in time”.  

Joskow (1991), page 70.

But can this approach explain why we regulate in the first place? Can it be said that the primary rationale for regulation is to protect the sunk investment of the monopoly firm?

If the primary purpose of regulation were to prevent expropriation of the monopolist’s sunk investment, this could be achieved through mechanisms which either:

(a) place a floor under the prices of the regulated firm; or
(b) commit the government to not interfering in the prices of the regulated firm.

If the primary purpose of regulation were protecting the investment of the monopolist, why are monopoly regulators routinely ostensibly concerned about high prices? If this hypothesis were correct, we would expect to see regulators routinely defending the regulated firm against actions by the government (or consumers) to lower prices. In practice, of course, most regulators seem more concerned with preventing excess monopoly rents rather than defending the monopolist against attempts by consumers or politicians to drive prices down.

Furthermore, this hypothesis does not explain why governments are tempted to interfere in the prices of these firms in the first place. After all, most OECD governments do not find it hard to commit to not interfering in the prices of most firms in the economy. What is it about monopoly firms which makes this temptation to interfere – to the point of threatening sunk investment – so much harder? Why can’t the government just commit to keeping its hands off?

surround utility operations. Finally, the fact that utility services tend to be massively consumed, and thus that the set of consumers closely approximates the set of voters, implies that politicians and interest groups will care about the level of utility pricing. Thus, massive consumption, economies of scale and sunk investments provide governments (either national or local) with the opportunity to behave opportunistically vis-à-vis the investing company.”

32 Joskow (1991), page 70.

33 This practice was not uncommon in the early days of regulation. Priest (1993) provides several examples where legislation placed an explicit lower bound on the price that could be set by regulation. “A gas franchise in Philadelphia in 1897, for example, set the price of gas at $1 per 1,000 cubic feet but provided that the rate could be changed by city ordinance. In order to protect the utility, the franchise prohibited the city council from reducing price below ninety cents prior to 1908; below eight-five cents prior to 1913; below eighty cents prior to 1918; or below seventy-five cents during the remaining ten years of the franchise” (page 310). An electricity franchise in Salt Lake City in 1893 “guaranteed the utility a minimum price per customer of $1.50 per month” (page 314).
Is it that the level of sunk investment required in regulated industries is larger than in other sectors of the economy? It is true that sunk costs are substantial in some regulated industries, but many large firms in the economy (such as car manufacturers, or aluminum producers) must also make a very substantial sunk investment. Most OECD governments do not seem to have too much trouble developing a reputation for not interfering in the pricing of these firms most of the time.

On the other hand, there are regulated industries where the sunk costs appear to be very low indeed. For example, until relatively recently, postal sorting was done manually, in a labour-intensive process, with very little sunk investment. Why is that regulation is not required to defend a cement mill from governmental opportunism, whereas regulation is required to protect a mail delivery company? Again, the hypothesis doesn’t seem to explain the facts.

Could it be that there is something special about the services sold by regulated industries which make it harder for governments to commit to not interfering in their pricing? For example, is it that regulated industries provide what we might call “essential” services. 34

Again, this does not seem to be an adequate explanation. There are many other goods and services (such as food, housing, energy) which are arguably even more “essential” for consumers than, say, postal services. Although there are occasional interventions in the prices of food or housing, these interventions are relatively rare and/or light-handed in developed economies – yet interventions in natural monopoly industries remain both widespread and carefully institutionalized (rather than ad hoc).

Spiller and Tommasi (2005) argue that one characteristic of a “public utility” is that the product or service is “massively consumed”. The suggestion is that regulation is required to protect the sunk investment of the monopolist when there are sunk costs and economies of scale and the product is consumed by a large proportion of the populace. On the other hand, the suggestion is that regulation is not required when there are sunk costs and economies of scale but the product is not widely consumed. But why then, has the Australian National Competition Commission recommended mandated access to an iron-ore-carrying rail infrastructure in the Pilbara region of Australia, when this service is only likely to be purchased by a handful of companies at the most? Most voters would never have even heard of this dispute. In any case, the product produced using this rail infrastructure (iron ore) is almost entirely exported, so Australian consumers are simply not affected. Again, this hypothesis simply does not fit the facts.

In summary, protecting the durable sunk investment of the regulated firm is certainly one significant factor that regulators must take into account when carrying out their task. But, I would argue, the significance of this factor is a consequence of the decision to impose price controls – it is not a driver for the existence of those price controls in the first place. Most monopolists could cover their sunk costs quite easily if they were simply unregulated (that is, allowed to charge whatever prices they liked). For these firms, allowing recovery of sunk costs is primarily about the government committing to keep its hands off. There are many ways that governments can commit themselves to keeping their hands off the assets of private firms – such as through constitutional prohibitions on “takings” and by developing a reputation for not expropriating sunk investments. These mechanisms seem to work adequately in most sectors of the economy. If the primary objective of regulation were the protection of the sunk investments of the regulated firm, regulators should be keenly focused on the risk of charging prices that are “too low”. In contrast, regulators seem to care just as much – if not more – about overcharging than they do about undercharging.

34 Bonbright (1988), page 8, lists as one of the characteristics of a public utility that it “provides a service that is ‘important’, ‘essential’, ‘vital’ – perhaps a ‘necessity’ for which present livelihood or future societal growth mandates the supply”.
There must be something else about these industries which requires a limitation on the pricing to end-users in the first place. In the presence of some other reason to limit the prices charged, the objective of protecting the sunk investment of the regulated firm becomes important, and can explain some of the features of regulatory practice. But what could be that reason for limiting the prices to end-users? This is the question we turn to now.
4. Natural monopoly regulation and sunk investment by end-users

In my view, a strong case can be made that the primary rationale for natural monopoly regulation is the protection and promotion of sunk investment – not the sunk investment of the monopolist, but the sunk investment of its customers and consumers. In other words, I claim that natural monopoly regulation is best viewed as a solution to a hold-up problem – but, in this context, it is the customers of the firm who are at risk of being “held up”.

The basic story is as follows: the users of the monopoly firm have the opportunity to take some action which incurs a sunk cost which will significantly increase the value of the monopolist’s product or services to them. The users or consumers, however, fear that once they have taken that action and incurred the associated sunk cost, the monopolist will engage in “ex post opportunism” - raising the price for the monopolist service, essentially expropriating the additional benefit or value achieved. Fearing this expropriation, the users or consumers are reluctant to put themselves in a position where they can be exploited by the monopolist. As a result, they fail to take socially efficient actions, or they take other actions which are less beneficial, but with lower risk of expropriation. The failure to take efficient actions results in a substantial economic welfare loss.

The monopolist realizes that its users and consumers fear being expropriated and tries to maintain incentives for customer investment through various mechanisms such as ex ante long-term agreements, developing a reputation for fair dealing, or directly incurring the sunk costs itself. However, these solutions are imperfect. In the long-run, in many industries, customer investment in sunk costs is best protected through the on-going oversight of a price-regulation authority who provides assurance to the customers that the monopolist’s prices will broadly stable, reflecting only changes in the long-run efficient costs of providing the services consumed by that user, and with no decline in service quality.

I attempt to show below that this hypothesis – that the primary rationale for natural monopoly regulation is the promotion of sunk customer investment – does a much better job of explaining regulators behaviour in practice.

What are the sunk actions taken by customers?

First, let’s look more closely at the precise nature of these sunk investments by consumers. There are many different actions that a customer might take which affects its demand for or value of the products/services of the monopoly. For example:

- A gas exploration company might be considering whether to invest in prospecting for gas in the vicinity of a single major gas pipeline. The cost of that investment is sunk. In the event that the prospecting is successful, the company knows that the gas will have to be carried by the monopoly pipeline owner. Secondarily, the same company might be considering whether or not to invest in R&D which could significantly enhance the effectiveness or efficiency of the gas exploration process.

- A worker might be considering whether to locate in the centre of a city, close to her place of work, or in a rural area, which would required substantially greater use of telecommunications services.

- A householder might be considering whether to install a long-lived electric hot-water relative to a more expensive system capable of heating water with both gas and electricity.

- A commuter might be considering whether to locate in a remote suburb, where she would be heavily reliant on the price and quality of the urban commuter rail network, as
opposed to an inner suburb, which would allow other commuting options such as walking or cycling.

- A company might be considering a substantial investment in R&D to develop a new technology which allows higher speed communications over copper pairs, say.
- A shipping company might be considering constructing a rail spur line from the interstate rail network to one of its terminals, so that it can promote and offer a door-to-door overnight delivery service to its own customers.
- A start-up airline might be considering offering budget flights from specific airports Australia to the South Pacific, which would require heavy up-front investment in promotion, and which would make it reliant on the maintenance of long-term reasonable pricing by those airports.

The literature on transactions costs, conventionally groups these different kinds of sunk investments into the following categories:

- The decision where to locate, when that decision will have an impact on the demand for monopoly services (e.g., close to a rail spur, close to a mine mouth, on which side of a river, in which suburb etc.). These are known in the transactions-costs literature as “site-specific investments”.

- The decision to invest in discovering, developing, or marketing a new product or service which makes use of the monopolist’s product or service as an input (strictly, as a complement) – e.g., the discovery of a new gas source, the discovery of a new telecommunications technology which allows higher speeds over copper infrastructure, the development of a new product. These are known as “human capital-specific investments”.

- The decision to invest in customer-premises equipment or other assets which are specialized to the monopolist’s product or service (such as telecommunications equipment, electrical equipment, gas consuming equipment). These are known as “physical asset-specific investments”.

Importantly, a customer can be said to make a sunk investment in reliance on the monopoly service not only by investing in a complementary specific asset, but also by not investing in a substitute asset. For example, a commuter increases his/her reliance on commuter rail services by not purchasing a car; a shipping company increases its reliance on rail services by selling off its fleet of specialized trucks; a manufacturer increases its reliance on local electricity supply by choosing not to purchase equipment capable of also burning natural gas. In these cases, the magnitude of the sunk investment depends on the extent to which the opportunity to purchase the asset at a given price is lost by failing to invest in advance (before the monopolist tries to increase the price).

This possibility of sunk costs being incurred on the demand side of the market has on occasion been recognised in the economics literature. For example, in 1976, Goldberg wrote:

\[ \text{See, for example, Crocker and Masten (1996), page 8.} \]
\[ \text{36 There is a clear parallel here to the notion of “avoidance costs” in Biggar (1995). This approach may also shed light on why consumers find some pricing practices “unfair”. For example, it is observed that consumers believe it is unfair for a local store to increase the price of snow shovels following a snow storm. (See Kahneman et al, 1986). This “unfairness” could be explained as a response by the consumer to the perceived inefficiency of numbers of consumers taking the “avoidance action” of holding an inventory of snow shovels in response to the possibility of hold up by the store when it is more efficient for that inventory to be held by the store.} \]
“To take advantage of a particular service might require the installation of long-lived complementary equipment. The consumer might be reluctant to purchase an all-electric house if he has no assurance that electricity will be available at ‘reasonable prices’ over the life of the house. On the other hand, the consumer will want to avoid being locked into an inferior technology: for example, he might want the flexibility to transmit data by microwave rather than telephone lines if he were to see fit”.

Gomez-Ibanez (2003) explains this in more detail:

“An effective monopoly in local infrastructure depends on the customers, as well as the company, making durable and immobile investments. The customers make their durable and immobile investments when they establish their residences and businesses in the territory served by the infrastructure company. These investments include the time a family must spend to find a suitable local home, job, and schools for the children, for example, or the resources a business devotes to developing a local workforce or customer base. The households and businesses could conceivably move to another community if their local infrastructure company increased prices. The costs of replacing their local and immobile investments are so high, however, that relocation is seldom a realistic option.

The infrastructure natural monopoly is a variant of the procurement problem in which relationship-specific investments are not optional but are an inevitable consequence of the technology. It is hard for an infrastructure company to avoid investments in durable and immobile facilities or for its customers to avoid investments specific to their communities. The company’s facilities cannot be transported to other locations if the local customers, or the government representing those customers, insist on price reductions. And the customers can’t easily relocate if the infrastructure company decides to raise prices”.

As far as I can tell, the implications of these sunk complementary investments by buyers for the theory of regulation has not been fully explored in the economics literature.

As long as the actions which buyers must take to increase their demand for or value of the monopolist’s service are sunk, buyers will fear that a proportion of the additional value created by these actions will be expropriated ex post. Anticipating this possibility, buyers will be reluctant to take actions which increase their exposure to opportunism by the monopolist. This may

37 Goldberg (1976), page 433.
38 Gomez-Ibanez (2003), page 9-10. There are a few mentions of the same idea in the wider economics literature (that is, outside the literature on regulation): Farrell and Gallini (1988), page 673, writing about switching costs, observed: “In many markets, buyers must bear specific setup costs in order to use a product. This can create a problem of opportunism: the seller can expropriate the returns to the buyer’s specific investment by raising the price ex post. … Buyers of a new product may be reluctant to incur setup costs if they will be exploited ex post”. Holmes (1990), page 789, writes: “Consumer investment in product-specific capital is a feature of the markets for many products, especially if one takes a broad perspective of what this capital decision can be. For instance, … the decision to reside far from work is analogous to the decision to buy a big car since (1) the decision may be influenced by the current [and future] price of gas and (2) the decision affects an individual’s future demand for gas. As another example, consider the demand for phone service. In response to a low price of phone service, businesses make capital decisions such as the purchase of computer telemarketing machines and other phone equipment. The businesses may also configure their marketing strategy to use phone contact rather than direct personal contact, and such a strategy involves investment in human capital. These investment decisions all tend to make the future demand for phone service relatively inelastic”.
39 The authors cited above, although they mention the possibility of relationship-specific investment on the buyer side, in practice focus almost exclusively on the implications of relationship-specific investment on the side of the regulated firm. Laffont and Tirole (2000), cited later, raise the implications of customer-side sunk investment for Ramsey pricing.
significantly reduce overall welfare. The hypothesis set out in this paper is that a primary rationale for regulation is to protect the sunk investment of buyers and therefore to promote on-going sunk investment by buyers in the present and the future.

Precisely this argument was made Charles River Associates in a recent submission to the Australian Productivity Commission. In their submission they argue against deregulation of Australia’s interstate rail infrastructure, precisely on the basis that it would allow the rail infrastructure owner to exploit complementary investments made by the above-rail companies, thereby having a “chilling effect” on any above rail investment:

“[R]ail will only develop as a viable transport option for North-South inter-modal freight if significant investment occurs in both below- and above-rail components. At the above-rail level, further increases in service quality, including through continuing innovation, are essential if rail is to be competitive. However, if the investments required to develop and implement these continuing improvements in service quality do occur, they will give rise to potentially substantial quasi-rents that will be vulnerable to expropriation.”

“The major concern with deregulating access prices for intermodal rail freight is that the infrastructure owner may take the opportunity to increase access prices to levels that would capture some or all of the above-rail operators’ return on and of capital (and other fixed costs). More specifically, the infrastructure owner would seek to shift to itself some of the quasi-rents associated with above-rail operators’ sunk investments. These include not only investments in physical assets, but also and very importantly, investments in expanding the use of the rail network, for example, by the development and marketing of innovative service options”.

Monopoly markets are not unique in requiring sunk investment by consumers to extract the full value of the product or service. In fact, in a wide variety of markets firms or consumers will make some degree of sunk investment whose value is contingent on continuing to obtain a supply of a good or service at a reasonable price. However, in competitive markets the value of those sunk investments is protected by the option of the buyers to find another supplier if the first buyer attempts to raise his price. It is precisely in those markets where buyers have few substitutes that any sunk investment they make is exposed to expropriation and therefore most likely to be deterred in the first place.

**Mechanisms for solving the hold-up problem**

We have observed that customers of natural monopoly firms will often have the opportunity to make sunk investments, giving rise to a hold up problem. As we will see in the next section, conventional natural monopoly regulation is one possible mechanism for protecting and promoting that sunk investment. But is it the only mechanism, or the best mechanism for protecting sunk investment? Doesn’t the monopolist itself have an incentive to design and implement mechanisms for promoting complementary investment?

Broadly there are several ways in which the monopolist could promote sunk investment:

(a) by reducing the cost of that investment to the buyer;

(b) by reducing the likelihood of hold-up;

---

40 CRA (2006), page 5. CRA note that the reliance on public subsidies increases the risk of hold-up: “The risk of hold-up issues arising in rail is particularly great. To begin with, below-rail operators currently depend heavily on subsidies. Any change in fiscal conditions that put downward pressure on subsidies would create pressures for hold-up to occur.” (page 5).

41 CRA (2006), page 3.
Vertical integration as a solution for the hold-up problem

The simplest mechanism for solving the hold-up problem is vertical integration between the monopolist and the buyer. By internalizing the costs and benefits of the sunk investment, the hold-up problem is eliminated.43

Historically, of course, the owners of many monopoly facilities were commonly vertically integrated into related sectors – even when those sectors were potentially competitive. For example, electricity companies were commonly integrated from the network monopoly to the related sectors of generation and retailing; gas pipeline companies were commonly integrated into upstream exploration and development and downstream retailing; telecommunications companies were, at one time, integrated into the provision of customer premises equipment and the manufacturing of telecommunications switches. The expansion of the monopolist into related sectors can be, in part, explained as a tool for the protection and promotion of sunk relationship-specific investment in the related sector.44

Vertical integration, however, introduces its own problems. First and foremost, it is not possible to vertically integrate with final consumers. To the extent that social efficiency requires sunk investment by final consumers, this approach cannot achieve the fully efficient outcome.

In addition, large firms may have governance problems of their own. It may be difficult to induce the necessary incentives for innovation. Individuals or firms who know that they have particular skills and innovations may find it difficult to communicate those abilities to the management of the firm, and therefore may have less incentives to acquire those abilities in the first place. Crocker and Masten (1996) explain:

"The benefits of integration are limited, however, by the bureaucratic inefficiencies that inevitably plague large organizations. Although, in principle, a newly integrated firm should be able to operate at least as efficiently as the two independent firms from which it was formed simply by allowing each division of the combined firm to operate independently as it had before and only intervening where net benefits were likely to be realized, the management of the combined enterprise will be unable, given the law governing internal transactions, to commit to intervening only in such a selected fashion. Without effective assurances that owners will not appropriate performance enhancements, the incentives of division managers to innovate, maintain assets, acquire and utilize information, and otherwise invest in the efficient operation of the division will be compromised. In their place, the firm is force to substitute weaker, indirect incentives dependent on managerial oversight. This attenuation of incentives combined with the limited capacity of management to administer additional transactions – which manifest themselves in a variety of bureaucratic inefficiencies – ultimately undermine the efficacy of internal organization and thereby limit firm size".45

42 There is a parallel here with the literature on switching costs: Switching costs are a form of sunk investment, which gives rise to the possibility of hold up or “supplier opportunism” which is partially protected by long-term contracts, competition combined with a commitment not to price discriminate, and reputation effects.
44 In a few industries, downstream users of the monopoly service jointly or collectively own the monopoly facility. Integration of this sort eliminates the hold-up problem. Such arrangements are common, in say, agriculture. In Australia, a group of carriers has proposed forming a joint venture to collectively own and operate a nationwide fiber-to-the-node network. This is another example of collective ownership of the natural monopoly.
In summary, vertical integration may partially restore the incentives for sunk investment by buyers when the monopoly product or service is used as an input into a downstream product or service. However, there remains a chilling effect on sunk investment by individuals or non-integrated downstream firms.

Other mechanisms for reducing the hold-up problem

There are various actions which the monopolist can take which enhance the credibility of its commitment to not expropriate any sunk investments by buyers, or which reduce the cost of making the sunk investment. These actions might include:

(a) second-sourcing / licensing the monopoly service to a third-party provider;

(b) most-favoured customer clauses, and/or other mechanisms for limiting the extent of price discrimination;

(c) direct funding of the sunk investment / leasing of the sunk asset.

Second-sourcing involves licensing the right to produce the monopoly product or service to an independent provider. If the licensing is for fixed terms and conditions over the long term, the buyer knows that it will have the option of obtaining the monopoly service elsewhere if the monopolist attempts to raise the price ex post. Second-sourcing is relatively common in some industries, but it is of limited usefulness in the case of natural monopoly industries, since – almost by definition, in natural monopoly industries, it is inefficient to duplicate the monopoly infrastructure.

An alternative approach is for the monopolist to essentially commit to not engaging in price discrimination – that is, promising to the buyer that he/she will receive a price at least as good as some other customer group. This approach works particularly well when there is a customer group whose demand is particularly sensitive to the monopoly price. In this case, by linking the two prices together (through a “most favoured customer” clause, for example) the monopolist can make a credible commitment to not raise the price to other “captive” customers.

Laffont and Tirole (2000) use as an example the case of an aluminium company which makes a substantial sunk investment reliant on a long-term supply of reasonably-priced electricity. Although the aluminium company may seek a long-term contract, it may also be possible to achieve the same objective through a non-discrimination clause:

“In practice the hold-up problem is … partly resolved through the external imposition of nondiscriminatory pricing. The aluminium company may not need to insist on a long-term contract if it knows that it will be able to buy electricity at the same price as other customers who are less dependent on electricity”.

For example, it may be that there is an on-going supply of “new” customers, who perhaps have not made a substantial sunk investment or who do not yet know the value of the monopoly service to them. The regulated firm may have an incentive to keep the price to these customers down, in order to encourage “penetration” of the monopoly service. As long as there remains a flow of “new” customers continues, other customers might be able to rely on this incentive to discount the prices offered to new customers, coupled with a non-discrimination clause, to provide assurance that any sunk investments they make will not be exploited. However, once the supply of “new” customers threatens to dry up, so will the protection of other customers and therefore the rate of sunk investment.

46 Laffont and Tirole (2000), page 74-75.
Another approach to promoting sunk complementary investment is direct subsidization or provision of the sunk investment by the monopolist. As already noted, in former years it was common for telecommunications companies to own and to lease customer-premises equipment.47

**Long-term contracts to solve the hold-up problem**

The final approach to solving the hold-up problem is through long-term contracts. The monopolist could, in principle, simply make a promise to “keep prices down” in the long-run, while maintaining service quality.48

We do, of course, see long-term contracts with monopoly service providers. For example, long-term take-or-pay contracts are common in the gas sector. Similar arrangements sometimes arise between airlines and airports, between water networks and water treatment facilities, or between above-rail operators and the operator of the interstate rail network.

But long-term contracts have their own problems. To begin with, negotiating a long-term contract is costly, so the transactions costs are higher, particularly when there are a large number of buyers.49 In the long-run, the costs and demand facing the monopolist may vary significantly, according to factors which cannot be foreseen at the time the contract was signed. It is impossible to negotiate and specify actions to be taken in every possible future contingency – long-term contracts are inevitably incomplete.

“Although parties will design contracts to balance the need for adaptation with the cost of effecting adjustments, the ability to define precise obligations in response to changing events in ways that can be enforced at low cost means that contracts will, on the one hand, tend to be inflexible and, on the other, leave considerable opportunity to cheat on the agreement or to attempt to evade performance”.50

In a complex or changing environment contractors rely increasingly on “relational” contracts, with the terms and conditions periodically adjusted by an independent third-party “arbitrator”:

“Transactors respond to the inability to write complete contracts in two ways. First, as the transaction becomes more complex or uncertain, contracts are likely to become more ‘relational’ in character. Rather than attempting to lay out a detailed specification of the terms of the agreement, relational contracts attempt to simply establish the process through which future terms of trade will be determined – ‘the establishment, in effect, of

---

47 It may also have been the case that early electricity companies subsidized consumer investment in electric appliances and lights.

48 In some industries there is a market in forward or future purchases of the service or commodity. These markets precisely allow the buyer to obtain a commitment to a fixed price in the future, eliminating the possibility of hold-up. However, it is precisely in those industries which are supplied by a monopolist that forward markets do not arise.

49 Gomez-Ibanez (2003), page 22: “Both the infrastructure company and its customers make durable and immobile investments that leave them vulnerable to opportunism. The logical solution is for the parties to sign contracts to protect themselves before they make their investments. … [H]owever, the transactions costs of negotiating and enforcing these contracts may be high, particularly if many small customers are involved or if their infrastructure requirements are complex and hard to predict”.

50 Crocker and Masten (1996), page 9. See also CRA (2006), page 5: “One option would be to lock in infrastructure charges through a very long-term contract. However, there are a number of obvious problems with this approach. To begin with, devising such a contract when the long term prospects for rail are so uncertain poses substantial difficulties. Additionally, the above-rail operator entering into such a contract would be exposed to the risk that better terms would be offered to its competitors in later periods, a risk that MFN clauses can help address, but not fully offset. Finally, the experience of above-rail operators is that it is difficult to secure such contracts on reasonable terms with government-owned entities.”
a constitution governing the ongoing relationship. Second parties will seek to reduce the costs of being bound to long-term agreements by adopting agreements of shorter duration.\textsuperscript{51}

Regulation as a solution to the hold-up problem

It has often been observed that natural monopoly regulation can be viewed as a form of long-term relational contract between the monopolist and his/her customers, with the regulator playing the role of the arbitrator mentioned above.\textsuperscript{52} According to the perspective put forward in this paper, a primary purpose of this long-term contract is to promote sunk investments by users and consumers. It does this by providing assurance to customers that price increases for each customer or each group of customers will be kept broadly in line with long-run efficient costs, while maintaining product quality in the long term.

According to this approach, regulation is the most efficient solution to the hold-up problem where economies of scale prevent the duplication of the monopoly facility, and where there are a large number of customers, so that the transactions costs of entering long-term individualized contracts is large, and where the monopolist cannot directly subsidise the buyer-side investments.

In summary, I have suggested in this section that a principal rationale for natural monopoly regulation is the protection of sunk investment on the part of users. It achieves this objective through protecting users (and sub-groups of users) against adverse price and quality changes in the long-term, thereby allowing users make sunk investments in reliance on expectations about future prices and quality. Let's turn now to exploring how well this approach can explain actual regulator behaviour.

---

\textsuperscript{51} Crocker and Masten (1996), page 9.

\textsuperscript{52} For example, Priest (1993): “For the public utilities and, I shall suggest, for other areas of regulation, the interaction between the regulator and the regulated firm or industry is difficult to distinguish from long-term contracting, dominated by predictable problems of unilateral or mutual adjustment over time in response to changing conditions”, page 294. Joskow (1991), page 66: “the set of regulatory rules and procedures that determine the prices that a regulated firm can charge are usefully conceptualized as a set of incentive or procurement contracts that link the regulator as a principal seeking to achieve some social or political objective and the regulated firm as the agent supplying goods and services that often require relationship-specific investments to support cost-minimizing exchange”.

5. Testing the sunk investment hypothesis

The previous sections put forward the hypothesis that the natural monopoly regulation is best understood as protecting sunk investments on the part of users. Does this hypothesis do a better job of explaining the regulatory patterns observed in practice than the deadweight loss hypothesis (or its alternatives) discussed above?

To answer this question, let’s draw out some of the implications or predictions of this hypothesis. Some of the implications of this hypothesis are:

1. Regulators will seek to preserve a degree of stability in services and in terms and conditions, even in the face of changing demand and supply conditions.

2. Regulation will focus more on preventing upwards shocks to prices than downward shocks and will focus more on preventing a decline in service quality than on ensuring an improvement in service quality.

3. Regulators will tolerate some forms of price discrimination but will actively discourage price discrimination which allows higher prices for customers which have made sunk investments.

4. Regulation will seek to ensure that revenues for any service or group of services cover the incremental cost of providing those services, in order to insulate the tariffs of groups of customers from the costs of serving other customers.

5. Regulation will be less likely to arise (and will be more “light-handed” when it does arise) when the sunk costs of customers are small and/or where customers have effective substitute sources of supply or where there are other solutions to protect sunk investments (conversely, regulation will be more likely to arise when sunk costs are large, when the risk of hold-up is large and when there are no other effective mechanisms for controlling hold-up).

6. Regulatory controls on prices may be necessary even when the regulated firm is earning a “normal” or below-normal rate of return.

Stability in services, terms and conditions

If, as hypothesized here, regulation has as a primary objective the protection of sunk complementary investments by customers, it follows that one of the primary concerns for regulators will be the promotion of long-run stability in regulated tariffs and service quality. Indeed, Bonbright (1988) includes in his list of the primary desirable attributes of good regulated tariffs:

“stability and predictability of the rates themselves, with a minimum of unexpected changes seriously adverse to rate-payers and with a sense of historical continuity”.

In fact, it seems to be conventional wisdom that regulators systematically seek to avoid sudden changes in tariffs, also known as “rate shock”. Kahn (1970) notes that:

“Growing public utility industries that are constantly adding to capacity generally must attempt to set their rates, as stably as possible”.


“The proper balance [between rate stability and cost reflexivity] will ordinarily have to make very large concessions in the direction of rate stability.”

Kahn cites Jackson (1968) as giving examples of electricity rates remaining absolutely unchanged for 10 and 15 year periods.

As noted earlier, in those regulatory regimes which allow a degree of pricing discretion to the regulated firm, it is common to find limitations on the rate of rebalancing (for example, in the form that no individual tariff can increase by more than, say, CPI+2%).

The reluctance to move quickly to more “efficient” price structures arises even when the existing tariffs are clearly below marginal cost. This is particularly clear in the case of the road transport industry where there are concerns in the EU about moving to more marginal-cost-based pricing.

As we saw in the first section of this paper, this reluctance to adjust tariffs or to limit the rate of rebalancing cannot be explained using conventional economic theory. According to conventional economic theory, limiting the rate of rebalancing merely reduces welfare by prolonging inefficient price structures. Although arguments based on risk aversion on the part of consumers may provide a partial explanation, conventional economic arguments do not provide a strong case for maintaining the stability of tariffs in the short or medium term.

In fact, not only do regulators tend to promote rate stability, they also tend to promote stability in the provision of services – even when those services are no longer strictly economic to provide. Kahn (1970) provides the example of the ICC which:

“refused to permit railroad abandonments of passenger service … on the basis of a variety of considerations: the national interest in preserving different transport media for possible emergency; a desire to maintain the quality of life in out-of-the-way communities; a feeling that the external benefits of continued passenger service to such communities may have justified the continuation even though private revenues fell short of private costs; recognition of the fact that railway property has traditionally been disproportionately heavily taxed”.

The sunk investment hypothesis provides one possible explanation as to why it might make sense to preserve services which are non-economic, even in the absence of externalities: a commitment to retain such services in operation for a period even when demand is insufficient ex post may be necessary to induce efficient investment ex ante.

Prevention of upward movement of prices or downward movement in quality

As already noted above, if the sunk investment hypothesis is correct, regulators would be expected to focus on providing a long-term stable path of prices and service quality. More precisely, we might expect that regulators would be particularly concerned not to allow increases in prices or decreases in quality.

The hypothesis that regulators will avoid increases in prices is difficult to test. Nevertheless there are some hints that this is, in fact, precisely what regulators seek to do. It was noted above that, according to Bonbright (1988), ideal rates should minimise “unexpected changes seriously adverse to rate-payers”, but makes no mention of avoiding changes to rates which are beneficial to rate-payers. Joskow (1974) emphasises that regulators have in practice acted in such a way as to prevent increases in prices:

“Contrary to the popular view, it does not appear that regulatory agencies have been concerned with regulating rates of return per se. The primary concern of regulatory commissions has been to keep nominal prices from increasing. Firms which can increase their earned rates of return without raising prices or by lowering prices ... have been permitted to earn virtually any rate of return they can. ... Consumer groups and their representatives (including politicians) tend to be content if the nominal prices they are charged for services are constant or falling.”

In addition, the provisions of Australia’s competition law which relate to “prices surveillance” (part VIIA of the Trade Practices Act) require notification of any increases in prices but do not prevent the regulated firm from selling at prices below the regulated level.

Similarly, there is at least some evidence that regulators focus more on penalizing a decline in standards than rewarding improvements in standards. In the Australian electricity industry, for example, transmission companies are required to meet statutory reliability standards, with relatively weak rewards for exceeding these standards.

In the US telecommunications industry, Lynch et al (1994) observe that the dominant approach at that time was to aggregate a number of performance dimensions into a single pass/fail decision. Companies which are held to have failed the performance standard were then punished in some way, with no offsetting rewards for exceeding the standard.

Broadly, the evidence suggests that regulators care more about preventing adverse movements in prices or quality than in promoting beneficial movements in prices or quality, consistent with the sunk investment hypothesis.

**Discouragement of forms of price discrimination**

If the sunk investment hypothesis is correct we would expect to see regulators preventing price discrimination when that price discrimination facilitates the exploitation of sunk complementary investments by consumers. The sunk investment hypothesis does not imply that regulators will be opposed to all forms of price discrimination – but rather only those forms of price discrimination which might raise the price charged to a customer after making the sunk investment. If the price faced by a particular customer is locked in – even if different from the price charged to other customers, the price discrimination does not threaten the sunk investment and, by this hypothesis, we would not expect to be opposed.

Let’s look first perfect price discrimination. As we saw above, according to the deadweight-loss hypothesis, perfect price discrimination is not only benign, it is welfare enhancing and therefore should be encouraged, in direct contrast to regulatory practice. However, in the light of the desire to protect sunk investment, perfect price discrimination would allow – and indeed would require – that the monopolist extract all of the additional rent created by any sunk actions on the part of the buyer which increased the buyer’s demand for or value of the monopolist product. In effect,

---

59 “The dominant approach to monitoring quality in regulated monopolies is to set performance standards on objective and technical dimensions of service quality (Model Telecommunications Service Rules, National Association of Regulatory Utility Commissioners 1987). The critical feature of regulation by standard is that essentially continuous variations in performance on any dimension are degraded into a two-category (pass/ fail) classification. … the Florida Public Service Commission has set standards on 38 rules governing local telephone companies. Public utility commissions in 30 other states in the U.S. employ similar systems of standards. … Regulators and companies collect data to determine whether the company has passed or failed the standard on each regulated dimension for an evaluation period.” Lynch et al (1994), page 175.
60 As noted above, we have seen that regulators (at least in the US) are almost always subject to the requirement that tariffs not be “unduly discriminatory”.
perfect price discrimination implies perfect expropriation of the sunk investment. If the sunk investment hypothesis is correct, perfect price discrimination is undesirable and should be prevented, consistent with observed regulatory practice.

Very similar arguments can be made about Ramsey pricing. Ramsey pricing requires a form of price discrimination – raising the price precisely on those users whose demand is more inelastic. It is highly likely that an action by a buyer which increased its demand for or value of the monopoly service, would imply an reduction in its elasticity of demand and therefore – according to Ramsey pricing – an increase in the price it is charged. In effect Ramsey pricing precisely allows, and indeed encourages, the monopolist to extract greater rents from those users who have made sunk investments.

Put another way, although Ramsey pricing is efficient in an economic framework in which buyers make no sunk investments, its efficiency does not carry over into a world in which buyers make sunk investments. In this world, in order to preserve incentives for investment, the monopolist must ensure that any price discrimination it carries out is independent of the actions of the buyer in making a complementary investment. This argument was mentioned by Laffont and Tirole (2000):

“Suppose that an aluminum producer builds a plant planning to use electricity rather than an alternative source of energy. Once the plant is built, the power utility can demand a very high price. Indeed, ex post Ramsey pricing implies that the utility fully extracts the aluminium producer’s profit (gross of the investment cost which is then sunk anyway). Anticipating this ‘special deal’ and knowing that it will lose the investment cost, the aluminium producer ex ante either does not build the plant or else selects its location and technology to fit a different source of energy, even though electricity may be the most cost-effective energy input. That is, the demand for electricity is more elastic ex ante than ex post. This example represents the familiar problem of expropriation of specific investments. The same problem is common in telecommunications. For example, a long-distance company or a value-added-service provider may be held up by the local loop provider after having made substantial investments”.

Peak-load pricing

Similar arguments may also explain the reluctance by regulators to wholeheartedly embrace peak-load pricing, while allowing some forms of time-of-use pricing.

Consider the position of a user considering making a sunk investment under a regime of peak-load pricing. That user would like to be able to predict the long-term path of prices it must pay in order to evaluate whether or not to make the investment. The long-term path of prices will depend on the peak/off-peak usage of that user and the peak/off-peak prices. The peak/off-peak prices depend, in turn on the long-run marginal cost of adding capacity to the monopoly facility, and the fraction of time the facility is operating at peak.

Importantly, under peak-load pricing, the regulator cannot provide an assurance to the user that prices in the future will not rise – either because demand is growing and the long-run marginal cost is upward sloping, or because demand is becoming more “peaky” (lower capacity ratio). As a result, the user may fail to make the sunk investment, even though it is socially efficient to do so.

This may explain why, when peak-load pricing is implemented, it is common to see incumbents and new entrant operators being granted physical or financial rights to the capacity of the monopoly facility (such as, for example, rights to the slots of a congested airport). Although

---

61 Laffont and Tirole (2000), page 74-75.
these rights might play a short-term risk-management role, the sunk investment hypothesis suggests a more important role – in facilitating efficient long-term sunk investment.\textsuperscript{62}

The role of incremental cost

One of the staples of regulatory practice has been the principle that the revenue obtained from providing a service or group of services, should at least be equal to the additional cost incurred in providing those services. This is sometimes known as the “incremental cost test”. Faulhaber (1975) points out that a version of this test “has been known in the public utility field for some time”, and cites an example from an 1887 book on railway rates by E. Porter Alexander.\textsuperscript{63} More generally, regulatory regimes around the world routinely require that tariffs reflect “costs” or “cost causality”, which is usually interpreted as implying that costs incremental cost as a price floor.

However, the basis in conventional economic theory for the requirement that tariffs exceed incremental costs is much weaker than is sometimes thought. As noted above, economic theory highlights the central role of marginal cost in pricing decisions. Deviations from marginal cost pricing depend upon the elasticity of demand for the service in question. According to standard economic theory, once it is determined that extending a natural monopoly facility to serve additional customers or provide additional capacity is socially efficient (that is, as long as the extra surplus received by consumers exceeds all of the additional fixed and variable costs associated with extending the facility), that extension should go ahead, regardless of whether or not the users of the additional capacity make a contribution to the fixed and common costs of that extension or not.

For example, consider the problem of pricing an expanding gas transmission network. Let's suppose that gas supply can be provided to an additional town at a cost of $10 million in fixed network costs and $1/GJ in marginal delivery costs. Cost-benefit analysis shows that extending the network in this way is welfare-improving overall. How should these costs be recovered? Economic theory is clear: the new users should pay at least the marginal cost of providing service (which is, in this case, $1/GJ). The allocation of the fixed costs depends only and entirely on the elasticity of demand of the new users and existing users. If the new users have elastic demand for gas and the existing users have an inelastic demand for gas, economic theory states that the additional fixed charges should fall partially or entirely on the existing users. The new users should pay as little as the marginal cost of the service.

This is sometimes expressed in the observation that Ramsey prices will not necessarily be “subsidy-free” in the sense that they will not necessarily cover incremental cost. Why then does regulatory practice routinely require that all users and groups of users pay the entire additional or incremental costs of the services they consume?

At this point, the objection might be made that the combinatorial cross-subsidy test proposed by Faulhaber (1975) requires that revenues from any service or group of services must exceed incremental cost – or else, the revenue from at least one service or group of services will exceed stand-alone cost. Since, this argument suggests, bypass of the natural monopoly facility is almost

\textsuperscript{62} Concern about the threat to buyer-side investment might explain the general difficulty observed in moving from “regional” to “nodal” wholesale markets in electricity. Even though move to more accurate or nodal pricing could be combined with the development of new short-term instruments for hedging the resulting trading risks, the sunk investment hypothesis suggests that market participants (generators and large loads) require much longer-term instruments (more akin to financial or physical rights to the capacity of the network) in order to justify long-term sunk investment in the presence of nodal pricing.

\textsuperscript{63} “If the [New York to San Francisco] price is more than the additional outlay in doing it, as against leaving it alone, it is profitable to the railroad, and the business is moreover advantageous to the whole inland community served by the railroad. For … the more prosperous the road, the lower the local rates may be made” E. Porter Alexander, \textit{Railway Practice}, New York, 1887, page 4, cited in Faulhaber (1975).
always inefficient, tariffs should be set so as to earn revenue below stand-alone cost. When combined with the condition that overall revenue must equal overall cost, this implies that revenue on any service or group of services must not fall below incremental cost. This, it is claimed, is the economic foundation for placing a floor on tariffs at incremental cost.

However, this result of Faulhaber only holds under certain conditions. Only in the extreme example of a pure “contestable” market, is entry a real prospect when revenue for a group of services exceeds stand-alone cost. In practice, in almost all industries, the threat of an immediate price response by the incumbent is sufficient to deter entry until revenues increase significantly above stand-alone cost – thereby allowing pricing on other services well below incremental cost. We could re-define stand-alone cost as the level of revenue at which entry is a real prospect, but this has the immediate effect of re-defining incremental cost as well. We can’t rely on Faulhaber (1975) to justify incremental cost – as conventionally defined – as a theoretically legitimate pricing floor.

But if the foundation of the incremental cost test in conventional economic theory is weak, why is it common in regulatory practice?

The sunk investment hypothesis provides a possible explanation. According to the sunk investment hypothesis, users and consumers would like some assurance of the long-term path of prices prior to sinking a complementary investment. If the monopoly facility is expanding, presumably the costs of providing that facility are also increasing. If there is a risk that that increase in costs will be reflected in increased tariffs for existing users, those users may be reluctant to sink any necessary investment, with consequences for economic efficiency.

According to this hypothesis, the regulator needs to give some assurance to users that they will not be forced to pay the costs of an expansion in the monopoly facility for which they do not directly benefit. One way to achieve this is to require that any additional fixed costs resulting from an expansion of the monopoly facility will be recovered entirely from the users of those additional services. A requirement that fixed costs are recovered from the users of the associated capacity ensures that individual users are insulated from changes in the network size which arise from the actions of others over which they have no control.

According to this hypothesis then, the primary rationale for the incremental cost test is that it insulates existing users from upward movement in tariffs brought about by investment to expand the monopoly facility which do not involve those existing users. Put another way, according to the sunk investment hypothesis, the incremental cost test is not about signaling to users the impact of their consumption decisions on the need to expand the network, or about preventing inefficient bypass of the network. Rather, the incremental cost test is designed to insulate existing users from increases in costs brought about by expansions in the monopoly facility which they do not require or use. A commitment to a rule such as the incremental cost test is one mechanism for providing some assurance to potential users as to the long-term stability of their tariffs.

**The Incidence of Regulation**

A key test for the sunk investment hypothesis is whether it can explain the incidence of regulation – that is, why certain services are regulated and others are not. The sunk investment hypothesis suggests that services are more likely to be regulated when:

(a) the customer must make material sunk investment;  
(b) there is a risk of hold-up (that the investment will be expropriated); and

---

64 Bonbright (1988), page 384, in his list of attributes of a desirable tariff structure expresses this as the absence of intercustomer burden. The full criterion is: “Avoidance of undue discrimination in rate relationships so as to be, if possible, compensatory (i.e., subsidy free with no intercustomer burdens)”. 

30
(c) there are no other effective mechanisms for controlling the hold-up problem.

It appears that these factors are all present in the “core” regulated public utility areas, such as electricity, telecommunications, local transport services, and so on. To test the hypothesis we must look to the “fringes” of conventional regulation.

For example, are there examples of services for which customers must make substantial sunk investments but for which the risk of hold-up is relatively small? Retail gasoline sales might be one such example. Households make substantial sunk investments in expectation of the future price of petrol in choosing say, the size or type of car to purchase, the length of commute, or even whether to purchase say, a holiday house accessible by car. Some businesses are also highly dependent on road transportation and sensitive to the price of road transportation relative to other transport modes.

The fact that households make substantial sunk investments dependent on the price of petrol may partially explain why petrol prices (and the possibility of collusion between petrol companies) remains a sensitive political issue in many countries (including Australia) and why retail prices for petrol in Australia have been periodically subject to various forms of regulation and control, despite the fact that the normal operation of the market mechanism provides at least partial protection of consumers against hold-up.\(^\text{65}\) If the sunk investment hypothesis is correct, the combination of a need for substantial sunk investment combined with a degree of competition places this market on the boundary between those market needing regulation and those which do not.

Another example where there is substantial sunk investment, but no regulation is the case of a coal-mine with a mine-mouth coal consumer, such as an electricity generator. If there are no other nearby sources of coal, choosing to site a generator at the mouth of the coal-mine involves a substantial sunk complementary investment. Yet, such investments are often protected through long-term contracts or vertical integration.\(^\text{66}\) No explicit regulation is required.\(^\text{67}\)

The sunk investment hypothesis can also explain why regulation of Internet domain name providers is a sensitive issue, despite the apparent lack of any economies of scale or natural monopoly considerations. The purchasers of Internet domain names often make substantial sunk investments in promoting their brand and website address – investments which are potentially subject to expropriation by the domain name provider.

Finally, the sunk investment hypothesis can explain aspects of compulsory licensing. Intellectual property rights, such as copyrights and patents, explicit create monopolies – to reward creative innovation. Why then, should those rewards be limited through compulsory licensing? One possible answer is that compulsory licensing prevents expropriation of sunk investments in complementary products which make use of the copyrighted or patented material. The agreement on Trade-Related Aspects of Intellectual Property Rights (“TRIPS”) allows for compulsory licensing to be imposed when “where the refusal to grant a license leads to the inability to exploit an important technological advance, or to exploit a further patent”.\(^\text{68}\) In other words, precisely in those cases where there is the greatest scope for expropriation of sunk investment.

\(^{65}\) Presumably larger consumers of gasoline could also purchase petrol in the forward market (perhaps linked to the price of a barrel of oil) to limit their exposure to the possibility of hold-up.

\(^{66}\) See Joskow (1985).

\(^{67}\) As another example, consider the case of “first-tier” tourist sites such as Disneyland in Anaheim. Tourists typically make a substantial sunk investment to travel to these sites, which is, in principle, subject to hold-up by the site operator, in the form of higher ticket prices for on-the-spot purchases. Yet, there are mechanisms for protecting such investments, such as advance ticket sales, by a commitment to no price-discrimination between advance and on-the-spot sales, or by developing a reputation for smooth changes in prices over time.

\(^{68}\) Source: Wikipedia.
The sunk investment hypothesis can also provide a clue as to why regulation has (historically at least) been focused on industries with a large number of customers—industries that Spiller and Tommasi (2005) refer to as being “massively consumed”. It is precisely in such industries that the transactions costs of negotiating long-term contracts with each consumer are prohibitive and for which conventional regulation has the clearest advantage.

**Regulation and subsidization**

Finally, the sunk investment hypothesis can explain why regulation may still be necessary for firms which do not even earn a normal rate of return. According to the sunk investment hypothesis the presence or absence of monopoly rents is not the primary driver of regulation—rather, it is the scope for hold-up. A firm may have significant scope to hold-up its customers even if it is earning below normal returns—and indeed, the customers of a firm may be particularly exposed to hold-up if that firm receives some external source of funding (subsidies) which can be withdrawn at any time.

The sunk investment hypothesis can explain why, for example, the Australian Competition and Consumer Commission (“ACCC”) is involved in regulating the terms and conditions of access to the interstate rail track network in Australia (operated by ARTC) despite the fact that ARTC would be earning below normal returns when its asset base is valued using conventional valuation techniques—the customers of ARTC must make substantial sunk investments in both physical and intangible assets and require significant long-term assurance as to the price and quality of service provided by ARTC. ARTC and its customers have benefited from direct government support in the past. The possibility that that support may be withdrawn may precisely act as a deterrent to making sunk complementary investment. The ACCC may be able to play a role in providing long-term price path assurance—again, despite the fact that ARTC is not earning any apparent monopoly rent.

**Other Implications**

The sunk investment hypothesis also highlights some likely key tensions in the work of the regulator—particular tensions between the objective of promoting long-term tariff stability, to promote sunk investment, on the one hand, and the variation in market demand and cost on the other. According to the sunk investment hypothesis, the regulator should seek to reproduce a hypothetical long-term contract between the regulated firm and its customers. Such a long-term contract would recognise the need to ensure sufficient revenues to cover long-term costs, as those costs evolve over time, while simultaneously promoting long-term stability in the prices to consumers. There is likely to be a key tension between allowing evolution in prices in response to changing supply and demand factors, on the one hand, and promoting price stability, on the other. To illustrate this, Bonbright (1988) provides some examples of problems caused by “promotional rates”:

> “The argument runs to the effect that the ratepayers were induced to locate their factors, or to abandon their isolated generating plants, or to convert their furnaces from coal to gas burners, in contemplation of the low, promotional rates and on the faith that this rate would remain in effect for the indefinite future. … [I]n the politics of utility rate regulation, the argument for stable rates is sometimes pressed with enough force to retard, for years, changes in rate structure otherwise clearly desirable. …

Many years ago, a number of manufactured-gas companies found themselves in this position after they had stimulated widespread resort to residential space heating by the establishment of heating rates originally justified because of redundant plant capacity. In railroad rate regulation, a striking example of a similar situation was that of the low commodity rates that the transcontinental railroads put into effect for eastbound Pacific Coast lumber, at a time when the balance of railroad tariff was heavily in the westbound
direction. … There was a long delay before the railroads succeeded in securing a corrective shift in their rate structure".  

The final observation we might make is that the sunk investment hypothesis can help explain the enduring popularity of what might be called the “competitive market benchmark” – the notion that regulators should seek to reproduce the outcomes of a competitive market. This view is, for example, expressed in Kahn (1970):

“the single most widely accepted rule for the governance of the regulated industries is regulate them in such a way as to produce the same results as would be produced by effective competition, if it were feasible.”

However, under the conventional economic approach, the competitive market benchmark provides virtually no practical guidance to a well-intentioned regulator. What aspect of the competitive market should a regulator seek to emulate? Pricing at marginal cost? But then how are fixed and common costs to be recovered? Pricing at average cost? But there is usually a wide range of combinations of prices consistent with full cost recovery – how should the regulator choose from within this range? Some of the key results from the economic of regulation – such as the benefits of price discrimination including Ramsey pricing and peak-load pricing either would not arise or could not arise in a competitive market. Despite its popularity as an aphorism, within the conventional economic framework the “competitive market benchmark” is simply of no practical help.

Why then does this notion persist? The sunk investment hypothesis provides a clue: in competitive markets consumers can be assured of the provision of services in the long-run at a price which broadly reflects costs and at a quality which reflects market demands. In addition, each customers tariffs are largely independent of the size of the supplying company and, in any case, do not depend on the consumption decisions of other users. These characteristics facilitate sunk investment. The competitive market benchmark is, in this view, a useful shorthand for the outcomes from regulation that consumers require.

Overall, it appears from this brief review that the sunk investment hypothesis can explain the broad patterns of the incidence of regulation, and in any case can do so better than the deadweight loss hypothesis. The sunk investment hypothesis allows us to explain the actual behaviour of regulators in a variety of circumstances, without recourse to arguments based on distributional concerns, equity concerns, or “political” concerns. The role for the minimization of deadweight loss seems to be limited at best. Rather, regulation appears to be about protecting consumers’ and users’ investments or innovations making use of the monopoly product.

Implications for competition law

Before concluding it is working noting the possibility that the sunk investment hypothesis may have implications for other aspects of competition policy – particularly competition law enforcement. As with the field of natural monopoly regulation, economists have long sought to understand and explain competition law enforcement in conventional economic terms. The conventional economic approach to competition law has focused – as with regulation – on the control of market power and the elimination of deadweight loss. Central to this analysis are concepts such as the Lerner Index (the price-marginal cost mark-up) and the elasticity of residual demand.

However I suggest that – as with regulation - the conventional economic theory of competition law enforcement has gained only limited acceptance by competition law enforcers. Despite the
efforts of economists in the last few decades, there remains something of a gap between the economic approach to competition law and the practice of competition law enforcers. This raises the question – does the sunk investment hypothesis put forward above allow us to better explain the actions of competition enforcers in practice?

For example, the conventional economic approach to antitrust matters would argue that only total economic welfare should be taken into account – that is, the sum of producers’ and consumers’ surplus. This argument was made, for example, in a well-known 1968 article by Oliver Williamson, in which he pointed out that even modest gains in productive efficiency could exceed the deadweight losses from price increases.\[72\]

But, at least in the US (and in many other important antitrust jurisdictions), mergers which would raise prices to users and consumers are not permitted even if there are demonstrable offsetting synergies or productive efficiencies. The basic rule in the US is that unless the productive efficiencies have the effect of lowering prices to consumers, the merger will be prohibited.

Economists have sometimes sought to rationalize this approach by arguing that the competition laws are designed to promote a consumer surplus or consumer welfare standard rather than a total welfare standard.\[73\] But this does not appear to be an adequate explanation either. After all, competition authorities regularly prosecute buyer cartels, or mergers to a monopsony which, if passed on, would result in lower rather than higher prices to end-users.\[74\] If promoting consumer welfare is the objective, why should competition authorities prohibit behaviour which may, in fact, lead to lower prices?

The obvious conclusion is that conventional economic theory does a poor job of explaining key aspects of competition law and the behaviour of competition law enforcers.

Does the sunk investment hypothesis do any better? According to the sunk investment hypothesis, the primary objective of competition law is to protect and thereby promote the sunk investment of buyers and sellers in the market. The sunk investment hypothesis, would justify blocking a merger if the merger allowed the merging parties to expropriate sunk investment by users – regardless of the size of the deadweight loss, and regardless of the magnitude of any productive efficiencies or synergies resulting from the merger. To this extent, therefore, the sunk investment hypothesis can better explain this core aspect of competition law.

It may be possible to push this argument still further. Why are all cartels conventionally treated as “per se” offenses rather than being analysed under a “rule of reason” approach. After all, some cartels may have economic efficiency benefits – either through more efficient production or through more efficient pricing. The conventional economic approach to competition law would argue that this is to economise on enforcement costs – to save the competition enforcer from having to demonstrate that the economic harm from the cartel exceeds the benefits. The sunk investment hypothesis suggests an alternative explanation – cartels allow the firms involved to better exploit the sunk investment of users and consumers and therefore should be prohibited regardless of their productive or allocative efficiency consequences.

More generally, the conventional economic approach to competition law focuses on the central role of concepts such as the price-cost margin, the Lerner index, and the elasticity of residual demand.\[75\] Yet these concepts are rare in competition law enforcement practice. Is this because

\[72\] Williamson (1968).

\[73\] For example, Kaplow and Shapiro (2007), page 87: “The modern trend in the United States seems to be toward a consumer welfare standard when considering the efficiencies defense”.

\[74\] Carlton (2007): “if only consumers matter, then a buying cartel should be perfectly legal and indeed should be encouraged”. Pittman (2007) argues that this argument only applies to a buying cartel formed by final consumers and does not apply to cartels of producers of intermediate goods.

\[75\] See for example Kaplow and Shapiro (2007).
these terms are difficult for lawyers and courts to handle? Or is it that competition law is better understood as being focused not on deadweight loss, but on protecting the sunk investment of users and consumers?

The sunk investment hypothesis may also explain the attitude of competition law and competition enforcers towards some forms of price discrimination. As noted earlier, despite the ambiguous welfare impacts of price discrimination according to conventional economic theory, competition law, at least in the US, has conventionally taken a dim view of price discrimination. Hal Varian in his chapter on price discrimination for the Handbook of Industrial Organization highlights this disjuncture between economic theory and legal reality:

“Price discrimination has long been regarded as a dubious practice from the legal viewpoint, though the complaints about the practice voiced by legislators are typically not those voiced by economists. …

In reading the legal discussion of the Robinson-Patman Act, one is struck by the difference between the legal concerns and the concerns of economists. The legal issues surrounding the issue of price discrimination and the Robinson-Patman Act are those of unfair competition, predatory pricing, and the like. The issues of concern to economists are those of efficient pricing.

… [W]e can expect that allowing price discrimination will typically enhance welfare if it provides a means of serving markets that the monopolist would otherwise not serve. Conversely, if the size of the market does not increase under price discrimination, there can be no net increase in consumers’ plus producers’ surplus. Thus, it would seem that an economically sound discussion of whether price discrimination is in the social interest should focus on the output effects. However, as we have seen above, this consideration has not played much of a role in the legal discussion of price discrimination”.

As noted earlier, the sunk investment hypothesis may give some insight into why price discrimination by firms with market power is discouraged: doing so protects and thereby promotes sunk investment by the users of the firm with market power.

Overall, the sunk investment hypothesis seems to offer promise at explaining key aspects of competition law and enforcement which have proved troubling for conventional economic theory.

Varian (1989), page 643-646.
6. Conclusion

This paper has an ambitious objective – to call into question the conventional economic rationale for natural monopoly regulation. In my view, the conventional economic rationale for regulation, as found in numerous economics textbooks, does a poor job of explaining the patterns of regulation we observe in practice. Is this because regulators are economically ignorant, or just not listening? Or is there something lacking in the conventional economic theory?

I have suggested in this paper that natural monopoly regulation can be better understood as designed to promote and protect sunk investments on the part of users and consumers of the monopoly product or service. Sunk investments are not unique to monopoly services – indeed, they will arise in most services – but it is only in the case of a monopoly that there is a risk of expropriation of the value of these investments.

I suggest that viewing natural monopoly regulation in this way helps us to understand the patterns and behaviour of regulators that we see in practice. For example, it helps to explain opposition to price discrimination, Ramsey pricing, and some forms of peak load pricing. All of these pricing approaches would allow – and indeed encourage – the monopoly firm to extract higher rents from customers which have made higher levels of sunk investments, and therefore would worsen the hold-up problem.

I suggest that the sunk investment hypothesis goes some way to bridging the still-broad gap between economic theory and regulatory practice.
References


Church, Jeffrey and Roger Ware, (2000), Industrial Organisation: A Strategic Approach, McGraw-Hill


Gomez-Ibanez, Jose A., (2003), Regulating Infrastructure: Monopoly, Contracts, and Discretion, Harvard


Kahn, Alfred, (1970), The Economics of Regulation: Principles and Institutions, MIT Press

37


