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

DARRYL BIGGAR*

Australian Competition and Consumer Commission and Australian Energy Regulator

Abstract

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. 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 explains 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.

* ACCC, GPO Box 520, Melbourne, VIC 3001, Australia. E-mail: darryl.biggar@stanfordalumni.org. The comments of an anonymous referee and participants at the ninth Annual ACCC Regulation Conference are gratefully acknowledged. The views expressed here are those of the author and are not necessarily the views of the ACCC or the AER.

1 Introduction

What is the primary public policy rationale for natural monopoly regulation? 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 let's now switch from a normative to a positive perspective. Does the hypothesis that natural monopoly regulation is primarily about minimizing deadweight loss do a good job of explaining the patterns of regulation we see in practice?

The answer is quite clearly no. In fact, there are many policy prescriptions which are soundly based in conventional "marginal" economic theory but which are partially or entirely rejected in regulatory practice, such as marginal cost pricing and Ramsey pricing. At the same time, there are patterns of regulatory practice which are hard to explain as an attempt to minimize deadweight-loss, such as the common emphasis on price stability, or the focus on incremental cost as a basis for pricing. Is this difference between theory and practice arising because economists are not making their voice heard? Or, is something missing from the conventional economic theory? Could it be that natural monopoly regulation is not primarily put in place to minimize deadweight loss?

I suggest that the conventional economic approach to natural monopoly regulation has neglected a key element of the picture. Rather than the minimization of deadweight loss, I suggest that natural monopoly regulation is often better understood as an attempt to protect 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 to 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 exist private mechanisms for controlling the hold-up problem, such as long-term contracts, these are not always feasible. In many cases, the best way to protect and promote sunk investments by users 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. But this literature focuses on promoting sunk investment by the regulated firm. Protecting the sunk investment of the regulated firm would usually imply placing a floor under the prices of the regulated firm – exactly the opposite of the behavior of regulators that is normally observed. Protecting the sunk investment of the regulated firm is a legitimate concern of the regulator once a decision has been made to impose regulatory controls, but this approach cannot explain the existence of the regulatory controls in the first place.

I suggest that the view that natural monopoly regulation is, in part, about protecting user investment, allows us to better understand and explain the patterns of regulation and the behaviour of regulators that we observe in practice. In particular, it allows us to explain why some of the key policy prescriptions of conventional economic theory are routinely ignored, without recourse to arguments based on "equity" or "fairness".[1]

In practice, natural monopoly regulation, like any public policy intervention, is the result of a political process and may seek to serve variety of ends. Economists have in the past suggested that in certain circumstances natural monopoly regulation can be understood as an attempt to favor one interest group over another or to redistribute income.[2] These alternative rationales for regulation are complementary to the approach set out here.

The approach taken here starts with the presumption that there is an underlying efficiency or welfare-maximizing objective of natural monopoly regulation, which is reflected, if only imperfectly, in the patterns of regulation we observe in practice. The aim of this paper is to identify that rationale for regulation which best fits with the observed regulatory practice. As we will see, there are significant weaknesses in the conventional efficiency-based or "market failure" approach to regulation. This paper suggests an alternative rationale which has been largely overlooked.

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 at other economic rationales for natural monopoly regulation, such 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 Can natural monopoly regulation be explained as an attempt to minimise deadweight loss?

Why do we regulate the prices charged by natural monopolies? 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 producing 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". According to conventional economic theory, the primary economic rationale for natural monopoly regulation is the minimization of that deadweight loss.[3]

[1] Concepts such as "fairness" or "equity", are potentially important, but are imprecise in practice. A wide range of regulatory decisions are potentially consistent with a "fairness" justification. My intention is not to dismiss the validity of these concepts but to propose a more precise, efficiency-based explanation for regulatory decisions which previously may have been justified under the heading of "fairness" or "equity"

[2] See, for example, Stigler (1971) or Peltzman (1976).

[3] The deadweight loss is also known as the "welfare triangle" or "Harberger triangle". See, for example, Braeutigam (1989, p. 1300) or Crocker and Masten (1996, p. 10). Hotelling (1938) suggests the notion of deadweight loss goes back beyond Marshall to Dupuit's work of 1844.

For the purposes of this paper, let's assume there is an underlying efficiency or welfare-maximization rationale for natural monopoly regulation, which is reflected, if only imperfectly, in the broad patterns of regulation we observe in practice. 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?

(1) Pricing at marginal cost

If the primary objective of regulation is the minimization of dead-weight loss, why do regulators place so little emphasis on ensuring that marginal prices approach marginal cost?

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[4], such as Kahn (1970), who emphasizes:

> 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.[5]

Nevertheless, despite the strong emphasis by economists, the relationship between marginal prices and marginal costs seems to be of secondary importance to regulators, at best:

> 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]

If natural monopoly regulation is primarily about minimizing deadweight loss, why has regulation in practice "deviated significantly from the first-best marginal cost ideal?"

(2) Price discrimination

Of course, one common objection to marginal-cost pricing is that, in the presence of economies of scale, a simple linear price equal to marginal cost would not allow the regulated firm to recover sufficient revenue to cover its total costs. But why insist on simple linear prices? Eliminating deadweight loss only requires that *marginal* prices be set equal to marginal cost. Why not allow various forms of price discrimination?

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. The marginal value (or marginal willingness-to-pay) of the last unit consumed is just equal to the marginal cost – eliminating the deadweight loss and maximizing allocative efficiency.[7]

---

[4] For example, Vickery (1955, p. 605) writes: "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". See also Joskow (1991, p. 69), Train (1991), and Viscusi, Vernon and Harrington (1995, p. 358-9).

[5] Kahn (1970, vol I, p. 65).

[6] Bonbright (1988, p. 415).

[7] See, for example, Train (1991, p.90) and Leeson and Sobel (2006). Leveque (2003) points out that perfect price discrimination also induces socially-efficient investment decisions: "In a nutshell … discrimination

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, no matter what the level of the monopolist's earnings.

But there is a consensus in the economics literature that perfect price discrimination would be uniformly rejected by regulators. Textbooks tend to mention perfect price discrimination primarily as a theoretical curiosity:

> The practical and ethical difficulties of primary price discrimination are formidable. Our purpose in describing price discrimination is not necessarily to recommend it as a form of regulation.[8]

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", by the standard theory.

The same argument applies to the use of a theoretically ideal two-part tariff. Suppose that a regulated firm came to the regulator with a proposal to use a two-part tariff, with the "fixed" part unregulated and varying between customers, and the "variable" part regulated and set equal to marginal cost. Let's suppose that the regulated firm could credibly demonstrate that it can set the fixed part of the tariff not so high as to deter some customers from consuming at all. According to the standard theory, the fully "efficient" outcome is achieved. But it seems likely that any such proposal would be rejected in regulatory practice.[9]

Conventional economic theory also has difficulty explaining legislative or regulatory controls on price discrimination more generally. According to conventional economic theory, price discrimination can be welfare improving – precisely when it reduces the deadweight loss associated with the exercise of market power. 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.[10] There are similar prohibitions on discrimination in the EU electricity regulations.[11] If price discrimination is economically efficient (according to conventional economic theory) why is it rejected or strictly circumscribed in practice?

---

based on observable characteristics of demand is able to achieve, or nearly achieve, both an optimal use of the existing network and an optimal level of investment".

[8] Train (1991, p.93).

[9] There are many other similar cases that can be constructed. For example, whenever the demand curve is inelastic at the "efficient" level of output, the monopolist can raise the price without increasing the deadweight loss. As long as a monopolist can show its pricing proposal does not distort the quantity consumed from the efficient level, it should be subject to no further regulation, no matter what the level of its "monopoly rents". Bonbright (1988, p. 103), uses this example to argue that there must be an "income redistribution" role for regulation.

[10] Bonbright (1988, p. 515): "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"..Joskow (1991, p. 69) suggests that rules against discrimination are designed to prevent regulators using their pricing powers to impose what amounts to implicit taxes and subsidies.

[11] Leveque (2003, p. 15): "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". Laffont and Tirole (1993), footnote 52 observe that the "rule of non-discrimination among consumers … has strangely been

(3) Ramsey pricing

In those cases where the regulator is unable to set the marginal price for each service equal to its marginal cost, economic theory still places central emphasis on reducing the deadweight loss. Specifically, according to conventional economic theory, the regulator should depart from marginal cost pricing in a way that minimizes the resulting deadweight loss. This implies, of course, Ramsey-Boiteux pricing.

Although there may be practical difficulties with implementing Ramsey pricing, within the conventional economic theory of regulation, Ramsey pricing is a full and complete solution to the pricing question.[12] Ramsey pricing ensures that the deadweight loss is minimised while ensuring that the monopolist earns no excess returns overall.

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. Laffont and Tirole (2000, p. 131) observe:

> "… it is fair to say that participants in the current regulatory debate are on the whole suspicious of Ramsey access pricing".

Despite the large number of regulatory decisions made each year, identifying decisions based explicitly on Ramsey principles is difficult.[13] 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.[14] One commentator sums up the status quo succinctly as follows: Ramsey-Boiteux pricing is "loved by economists but spurned by regulators".[15]

But why, exactly? Textbooks suggest various reasons why regulators are 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.[16] But, 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, than it is to determine, say, whether or not a major investment project should be allowed to proceed?[17]

---

interpreted by charging identical prices to consumers with vastly different marginal cost of service (e.g., city and rural customers)".

[12] Faulhaber and Baumol (1988, p. 595) 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".

[13] For example, Decker (2007) comments that "… regulatory agencies have typically eschewed Ramsey-based pricing approaches in practice… " and have shown a "general reluctance" to implement Ramsey pricing.

[14] See Decker (2007, p. 10).

[15] Albon, Rob, personal communication, 20 August 2007.

[16] 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".

[17] Decker (2007, p. 13) raises the question whether the information requirements for Ramsey pricing are much larger than the other tasks required of the regulator...

Textbooks also routinely point out that Ramsey pricing may not be adopted due to distributional or equity concerns.[18] A full discussion of these issues is beyond the scope of this paper, but we can observe that there must be something missing from the central economic theory. The hypothesis that the primary rationale for regulation is the elimination of dead-weight loss cannot explain the widespread resistance to Ramsey pricing observed in practice.

(4) Incremental cost

We have seen that the deadweight-loss hypothesis has a hard time explaining why regulators fail to pursue policies which are efficient under the conventional economic theory, such as Ramsey pricing. In addition, the deadweight-loss hypothesis has difficulty explaining why regulators actively pursue policies which have little or no justification under the conventional theory, such as the focus on ensuring price stability, or the emphasis on incremental cost as a pricing floor.

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"[19]. Regulatory regimes around the world routinely require that tariffs be "cost based" or "cost reflective" – terms which are usually interpreted as implying incremental cost as a price floor.

Economists have accepted this emphasis on incremental cost largely uncritically. But the basis in conventional economic theory for the requirement that tariffs be based on incremental costs is much weaker than is sometimes thought.

As noted above, rather than incremental cost, economic theory highlights the central role of *marginal cost* in pricing decisions. Ramsey prices – which are optimal under the conventional theory under a wide range of circumstances – may indeed be lower than incremental cost.[20] This is often summarized in the observation that Ramsey prices will not necessarily be "subsidy-free".[21] 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?

The seminal paper by Faulhaber (1975) is often cited as the basis for pricing above incremental cost. Faulhaber observes that since bypass of the natural monopoly facility is almost always inefficient, tariffs on any service or group of services should be set so as to earn revenue at or below the replacement 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.

However, this result of Faulhaber only holds under very narrow assumptions. Only in the extreme example of a pure "contestable" market, is entry a real prospect when revenue for a group of services exceeds the replacement cost of providing those services. In

[18] For example, Laffont and Tirole (2000, p. 132) note that most experts oppose Ramsey access pricing on the grounds that it does not satisfy the requirement of being "fair and non-discriminatory".
[19] Faulhaber (1975) cites E. Porter Alexander, *Railway Practice*, New York, 1887, page 4.
[20] For example, let's suppose it is socially efficient to expand an existing network to service new customers. If at least one set of existing customers have highly inelastic demand, Ramsey pricing will require that the fixed costs of the network expansion fall entirely on those existing customers, with service to the new customers provided at marginal cost.
[21] See, for example, Faulhaber (1975, p. 973), or Church and Ware (2000, p. 798).

practice, in almost all industries, if the entrant duplicates the facilities of the incumbent it will at best be able to capture a portion of the existing market, and so, in the presence of increasing returns, will incur higher average costs. In addition, in most cases the threat of an immediate price response by the incumbent is sufficient to deter entry until revenues increase significantly above replacement cost. Although there may arise some limited scope for bypass or duplication of the existing network by entrants which are particularly well placed (such as an electricity generator located next door to a large electricity consumer), in most natural monopoly industries, any significant bypass would likely require a reasonable prospect of revenues much in excess of the replacement cost of the existing network and may simply not occur at all.

Where pricing above replacement cost implies no immediate risk of bypass, there is no justification for insisting on pricing above incremental cost. We simply cannot rely on Faulhaber (1975) to justify incremental cost as a theoretically legitimate pricing floor across all industries.[22] So why, then, do regulators place so much emphasis on incremental cost?

(5) Price / service stability

Another puzzle for the conventional economic approach to regulation is the heavy emphasis on price stability. There is a sizeable amount of evidence that price and service stability is one of the primary concerns of regulators. According to Bonbright (1988) one of the primary desirable attributes of good regulated tariffs is:

> 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.[23]

Similarly, Kahn (1970):

> Growing public utility industries that are constantly adding to capacity generally must attempt to set their rates as stably as possible.[24]

Bonbright (1988) observes:

> [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. …[25]

This focus on long-term rate stability is such that even 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 of prices (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

[22] There is one possible argument for incremental cost as a price floor which, to my knowledge, has not been mentioned in the regulation literature. Following Coase (1946), this argument states that the requirement that incremental revenues exceed incremental cost ensures that sufficient surplus is generated from each service or project to justify it being carried out at all. In the absence of this requirement, it could be argued, the regulator would have to rely on imperfect assessments of demand and consumers' surplus to assess whether or not a project or new service should be allowed to proceed. The role of the incremental cost floor is to alleviate the risk that projects will proceed where the net social benefit is negative.

[23] Bonbright (1988, p. 383).

[24] Kahn (1970, vol I, p. 107-108).

[25] Bonbright (1988, p.188). Similarly, a central thesis of Owen and Braeutigam (1978) is that a major purpose of regulation is to protect economic agents (consumers and firms alike) from too-sudden changes to their economic environment.

cost. For example, there have been clear concerns about moving to more marginal-cost pricing in the road transport industry in the EU.[26]

The insistence on stability is hard to explain under the conventional economic approach. After all, if a price is inefficient, is it not preferable to eliminate the inefficiency as rapidly as possible?

Not only do regulators tend to promote rate stability, they also tend to promote stability in the set of services provided – even when some of those services are no longer strictly economic to provide. For example, Kahn (1970) cites the ICC which "refused to permit railroad abandonments of passenger service"[27]

This heavy emphasis on price and service stability simply makes no sense under the conventional economic approach.

(6) Prevention of upward movement of prices or downward movement in quality

Moreover, regulators are not just interested in promoting stability in prices - regulators seem to be routinely more concerned about preventing increases in prices than preventing decreases.

There is reasonable evidence that regulators care more about changes in prices and quality which are adverse to users than changes beneficial to users. Under the conventional approach in the US, public utilities are required to file a "rate case" in order to increase their prices, but may choose not to do so (essentially allowing their prices to fall in real terms) indefinitely. In Australia, the provisions of Australia's competition law which relate to "prices surveillance" (part VIIA of the Trade Practices Act) require notification of any increase in prices but do not prevent the regulated firm from selling at prices below the pre-approved level.

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.[28]

Overall, there is a strong suggestion that regulators focus more on preventing increases rather than decreases in regulated rates. This asymmetric attention paid by regulators to changes in prices is difficult to explain under the deadweight-loss hypothesis – after all, if prices are set so as to eliminate the deadweight loss, a reduction in the price (relative to marginal cost) is just as likely to harm welfare as an increase in the price.

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

[26] See Milne, Niskanen and Verhoef, (2000).
[27] Kahn (1970, vol I, p. 192).
[28] Joskow (1974, p. 297-298).

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. [29]

Overall, as we can see, the hypothesis that regulation is primarily about minimizing deadweight loss does a relatively poor job of explaining the patterns of regulation we observe in practice. This is not to imply that there may arise instances where regulators seem to have a clear focus on deadweight loss. However, those instances don't seem to be part of the broad observed patterns of regulation. Regulators must have other concerns in mind when they make their decisions. But what might be this alternative rationale for regulation?

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.

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, is regulation necessary to maintain or promote productive efficiency? The normal governance mechanisms on 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".[30]

One possible reason why productive efficiency pressures may be weaker on monopoly firms is due to a lack of suitable similar comparator firms. The presence of comparator firms allows external owners and shareholders to better identify common factors affecting the performance of all firms and thereby to better isolate an effective signal of the performance of the management of the firm in question. In the absence of good comparators, 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 an

[29] Lynch et al (1994, p. 175).
[30] See, for example, Church and Ware (2000, p. 145).

absence of suitable comparator firms is the underlying problem, why does regulation persist even when potential comparator firms are present? Why does it remain common, for example, to regulate electricity distribution businesses of which there are often dozens, even in relatively small countries?

Furthermore, if the promotion of productive efficiency is the objective, is natural monopoly regulation the right tool? After all, many economists would argue that natural monopoly regulation hinders rather than promotes incentives for productive efficiency.[31]

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 as in the case of 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. All of the monopoly rent would then accrue to the government, allowing it to reduce other distortionary taxes while eliminating the 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.

3.1 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 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.[32]

Similarly, 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.[33]

This strand of the economics literature is important and has enhanced the understanding of certain key issues in the design of regulatory institutions. 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:

[31] See for example Church and Ware (2000, p. 847) and the discussion of the Averch-Johnson effect.
[32] Joskow (1991, p. 67-68). See also Crocker and Masten (1996, p. 30).
[33] Spiller and Tommasi (2005, p. 5), emphasis added. See also Gomez-Ibanez (2003).

> 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.[34]

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 place a *floor* under the prices of the regulated firm[35] or which commit the government to not interfering in the prices of the regulated firm. If the primary purpose of regulation were protecting the sunk investment of the monopolist, why are monopoly regulators 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 (at least in OECD countries) seem more concerned with preventing excess monopoly rents rather than defending the monopolist against attempts by consumers or politicians to drive prices down.[36]

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. 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", or by developing a reputation for not expropriating sunk investments. These mechanisms seem to work adequately well in most sectors of the economy. What is it about monopoly firms which makes this temptation to interfere – to the point of threatening sunk investment – so much harder?

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

---

[34] Joskow (1991, p. 70).

[35] This practice was not uncommon in the early days of regulation. Priest (1993, p. 310) 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". An electricity franchise in Salt Lake City in 1893 "guaranteed the utility a minimum price per customer of $1.50 per month" (p. 314).

[36] Regulators do, on occasion, defend the regulated firm against the opportunistic behaviour of governments. Andres, Guasch and Straub (2007, p. 47) in a study of concession contracts in Latin America note that "in the case of government-led renegotiation, the regulator acts as a barrier against political opportunism. Regulation attempts to protect investors and ultimately consumers from the opportunistic behaviour of the government".

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 it 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 fit 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"? [37]

Again, this does not seem to be an adequate explanation. There are many other goods and services (such as food, housing, or fuels) 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. But why then, does regulation seem to persist even when the monopoly service is used exclusively in exporting industries (such as a natural gas pipeline or rail line serving export markets)?[38] Again, this hypothesis does not seem to 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. Many monopolists could cover their sunk costs quite easily if they were simply unregulated and allowed to charge whatever prices they liked. There must be something else about these industries which requires a limitation on the pricing to end-users in the first place. 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 case can be made that a key component of the 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.

[37] Bonbright (1988, p. 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".

[38] For example, the Australian National Competition Commission ("NCC") recently recommended mandated access to an iron-ore-carrying rail infrastructure in the Pilbara region of Australia. See NCC, "Fortescue Metals Group Ltd", Final Recommendation, 23 March 2006.

The basic story is as follows: the users of a monopoly firm routinely have the opportunity to take some irreversible action which will significantly increase the value of or demand for the monopolist's product or services. 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, 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 socially beneficial, but with lower risk of expropriation. The failure to take efficient complementary actions results in a material economic welfare loss.

The monopolist, which realizes that its users and consumers fear being expropriated, might try 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 sunk investment is best protected through the on-going oversight of a price-regulation authority who provides assurance to the customers that the monopolist's services and service quality will be maintained and that prices will broadly stable, reflecting only changes in the long-run efficient costs of providing the services consumed by that user.

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 the way that regulators behave in practice.

4.1 What are the sunk actions taken by customers?

There are many different irreversible 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. In the event that the prospecting is successful, the company will be forced to rely on the transportation services of the monopoly pipeline.[39]

- 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, requiring greater reliance on the monopoly provider of telecommunications services.

- A householder might be considering whether to install a long-lived electric hot-water system, increasing its reliance on the local monopoly electric utility, or whether to purchase a more expensive dual-fuel 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.

[39] 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 technology research company might be considering a substantial investment in R&D to develop a new technology which allows higher speed communications over copper pairs. If successful, the company would require access to the incumbent monopoly copper network.

- A shipping company might be considering constructing a rail spur line from the interstate rail network to one of its terminals, increasing the range and quality of services it can offer to its own customers, but also increasing its reliance on long-term access to the monopoly interstate track.

- A start-up airline might be considering offering budget flights from specific airports, 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"[40].

- 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 have been categorized 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".[41]

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, Gomez-Ibanez (2003) explains:

> 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

[40] See, for example, Crocker and Masten (1996, p. 8).

[41] Importantly, a customer can be said to make a sunk investment in reliance on the monopoly service not only by investing in a sunk complementary specific asset, but also by *not investing* in a substitute asset when the decision to pass up an opportunity to purchase a substitute asset is at least partially irreversible. For example, a shipping company increases its reliance on rail services by selling off its fleet of specialized trucks (which it cannot buy back at a reasonable price later); a manufacturer increases its reliance on local electricity supply by choosing not to purchase equipment capable of also burning natural gas. These actions are irreversible if the opportunity to purchase the substitute asset at the current price is temporary. There is a clear parallel here to the notion of "avoidance costs" in Biggar (1995).

> 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.[42]

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.[43]

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 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.

For example, a recent submission to the Australian Productivity Commission argues against deregulation of Australia's interstate rail infrastructure – even though there is no evidence of any "monopoly rents" – 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 above-rail investment:

> 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.[44]

---

[42] Gomez-Ibanez (2003, p. 9-10). See also Goldberg (1976, p. 433). The same notion occasionally arises in the wider economics literature (that is, outside the literature on regulation): Farrell and Gallini (1988, p. 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, p. 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".

[43] Laffont and Tirole (2000), cited later, raise the implications of customer-side sunk investment for Ramsey pricing. Owen and Braeutigam (1978, p. 35-36) recognise the weaknesses of the deadweight–loss hypothesis, noting that "It is easy enough for economists to get caught up in the fantasy of assuming the world cares or should care about some narrowly defined notion of efficiency". They recognise there could be other objectives of regulation: "It becomes at least arguable that regulation, at the cost of some efficiency and of some progressivity, may have provided substantial benefits to individuals by protecting them from some of the risk they would otherwise face from the operation of the efficient but ruthless free market". But they focus on the process of regulation, rather than explicitly on the sunk complementary investments by consumers.

[44] CRA (2006, p. 3).

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 supplier 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.

4.2 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?

There are several means by which a monopolist might seek to promote sunk investment by its consumers:

- by reducing the cost of that investment to the buyer;
- by reducing the likelihood of hold-up;
- by entering the market of the buyer and thereby internalizing the risk of hold-up.[45]

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.[46]

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.[47]

---

[45] 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.

[46] See Crocker and Masten (1996, p. 9).

[47] 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.

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, the transactions-costs literature emphasizes that large firms may have governance problems of their own. Crocker and Masten (1996, p. 9) observe:

> 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.

Besides vertical integration, there are various actions which the monopolist might take to enhance the credibility of its commitment to not expropriate any sunk investments by buyers, or which reduce the cost of making the sunk investment, such as:

- second-sourcing / licensing the monopoly service to a third-party provider;
- most-favoured customer clauses, and/or other mechanisms for limiting the extent of price discrimination;
- direct funding of the sunk investment or 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, by definition, it is inefficient to duplicate the monopoly infrastructure.

An alternative approach is for the monopolist to commit itself to not engage in price discrimination by promising to the buyer that he/she will receive a price at least as good as some other customer group (through a "most favored customer" clause, for example). This approach is particularly effective when there is a customer group whose demand is sensitive to the monopoly price. In this case, by committing to not charge any more than it charges to these "footloose" customers, the monopolist can make a credible commitment to not raise its price to the "captive" customers.[48]

Laffont and Tirole (2000, p. 74-75) use the example of an aluminum 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, allowing the aluminum company to purchase electricity at the same price as other customers who are less dependent on electricity.

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

[48] Holmes (1990) works through a fully-specified model in which consumers must make a decision as to the size of car to purchase, and a monopolist sets the price of gasoline. There is a constant supply of new customers, and the monopolist is unable to discriminate between customers on the basis of the size of car.

was common for telecommunications companies to own and to lease customer-premises equipment.[49]

Finally, a monopolist might seek to prevent the hold-up problem by committing itself through a long-term contract. The monopolist could, in principle, simply make a contractual promise to "keep prices down" in the long-run, while maintaining service quality.[50] We do, of course, observe long-term contracts between buyers and monopoly service providers in some industries. 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 rail track owners.

But long-term contracts have their own problems. To begin with, negotiating a long-term contract is costly, so the transactions costs are high, particularly when there are a large number of buyers.[51] 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.[52]

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 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.[53]

This brings to us the role for regulation. 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.[54] According to the perspective put forward in this paper, a primary

---

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

[50] 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.

[51] Gomez-Ibanez (2003, p. 22): "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".

[52] Crocker and Masten (1996, p. 9): "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". See also CRA (2006, p. 5).

[53] Crocker and Masten (1996, p. 9).

[54] For example, Priest (1993, p. 294): "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". Joskow (1991, p. 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

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.

5 Testing the sunk investment hypothesis

The previous sections put forward the hypothesis that natural monopoly regulation is best understood as a mechanism for 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?

In fact, the sunk investment hypothesis is consistent with the observed emphasis on price and service stability, the emphasis on preventing adverse movements in prices and service quality, the aversion to Ramsey pricing and price discrimination, and the emphasis on incremental cost as a pricing floor.

If, as hypothesized here, regulation has as a primary objective the protection of sunk complementary investments by customers, it follows that regulators will be concerned to maintain a stable path of prices over time since, in the absence of "forward" or "future" prices for regulated services, uncertainty over future tariffs or service quality will have a chilling effect on sunk investment by users. [55]

Furthermore, the sunk investment hypothesis is consistent with regulators' desire to promote the stability of the set of services provided, even to the extent that some services may be provided which are non-economic: in the presence of uncertain future demand, a commitment to retain services in operation for a period even when demand is insufficient ex post may be necessary to induce efficient investment ex ante.[56]

The sunk investment hypothesis is also consistent with the observed asymmetric attention paid by regulators to adverse movements in prices or service quality. If regulators are primarily concerned with protecting and promoting sunk investment we might expect that regulators would be particularly concerned not to allow increases in prices or decreases in service quality since it is only the risk of "bad news" outcomes which threatens ex ante investment.[57]

---

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".

[55] Long-run stability (or at least predictability) of the price-path requires either a flat long run marginal cost curve and/or predictability of long-run demand. If the long-run marginal cost curve is upward-sloping and long-run demand is uncertain, the regulator cannot guarantee any long-run price path. In this case, it may be preferable – from the perspective of promoting sunk investment - for the regulated firm to sell long-term capacity access rights. Rights to take-off and landing slots at airports is one such example.

[56] There are hints of the role of sunk costs in promoting price stability in Owen and Braeutigam (1978, p. 20-21) who observe that "A very primitive, minimum response to [the desire to protect victims of economic change] is the grant of a period during which adjustment can take place and useless fixed costs amortized. Non-economists are great respecters of sunk costs; the transformation of useful physical and human capital into an irrelevant sunk cost by a market or technological forces is a process that is easily viewed as unjust and even inhumane". I argue that the stranding of sunk investment by consumers is viewed as "unjust" or "inhumane" because it is inefficient, or more precisely, the risk of such stranding deters efficient investment.

[57] As in the Bernanke "bad news" principle.

If the objective of regulation is the promotion of sunk investment by consumers, certain forms of price discrimination could be undesirable. Specifically, if price discrimination were allowed the monopolist might seek to raise the price on those users who have made the largest sunk investment. Preventing this form of price discrimination, prevents expropriation of the value of that sunk investment, and thereby promotes efficient sunk complementary investment.

As we saw above, according to the deadweight-loss hypothesis, perfect price discrimination is not only benign, it is welfare enhancing and should be encouraged. However, perfect price discrimination would allow – and indeed would *require* – the monopolist to extract all of the additional rent created the user's sunk actions. In effect, *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.

Similar arguments can be made about Ramsey pricing. Laffont and Tirole (2000) explain:

> 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.[58]

If fact, Ramsey pricing does not always allow the regulated firm to expropriate the value of a consumer's sunk investment – that depends on the impact of sunk investment on the elasticity of demand of the consumer.[59] However, from the regulator's perspective there is a clear risk that Ramsey pricing will allow, and indeed encourage, the monopolist to extract greater rents from those users who have made sunk investments. Ramsey pricing, although effective at reducing dead-weight loss, may be incompatible with the promotion of sunk complementary investment.

Finally, let's examine the possible rationale for incremental cost as a pricing floor. We saw earlier that, under the dead-weight loss hypothesis, the economic foundation for using incremental cost as the basis for regulated tariffs is weak or non-existent. Why, then, does incremental cost receive so much attention from regulators?

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. One of the risks that

---

[58] Laffont and Tirole (2000, p. 74-75).

[59] For example, suppose that one hundred consumers have identical demand for a service which is inelastic up to a choke price P. Each consumer only has demand for a single unit of service. Suppose the monopolist facility costs $1000. As long as the choke price is above $10, the Ramsey price for each customer can be set at $10 per unit. If the effect of the consumer-side sunk investment is only to significantly increase the choke price for those customers, the Ramsey-optimal price can remain unchanged for all customers even if discrimination between types of customers is allowed – there is no expropriation of the value of the sunk investment. If the effect of the sunk investment is also to increase the quantity demanded, the Ramsey-optimal price may, in fact, be reduced for all customers.

consumers might face is the risk that, after making a sunk investment, the monopoly facility will be expanded to service other new customers. If existing users will be required to bear some of the costs of that expansion of the network, in the form of higher prices, those users may be reluctant to sink any necessary complementary investment.

According to this view, 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.[60] One way to achieve this is to require that any additional costs arising from an expansion of the monopoly facility will be recovered entirely from the users of the additional services. A requirement that all the incremental costs of an expansion are recovered from revenues received from the incremental users of the expansion ensures that existing users are insulated from changes in the network size over which they have no control.

Although the central role of incremental cost is difficult to understand if regulation is primarily seeking to minimize deadweight loss, it makes sense as a mechanism for providing some assurance to potential users as to the long-term stability of their tariffs.

The sunk investment hypothesis suggests that price regulation is most likely to arise where the customer must make a material sunk investment in reliance on a service, where that investment is exposed to the risk of an increase in the price for the service, and where there are no other effective mechanisms for controlling the hold-up problem. This hypothesis may explain why, say, gasoline prices remain a sensitive issue in many countries, since households must make a significant sunk investment in reliance on transportation, or why regulation of Internet domain name providers is common despite the lack of any sunk investment on the part of the providers, since purchasers of domain names must often make substantial sunk investments in promoting their brand and website address.

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 (such as government subsidies) which can be withdrawn at any time.[61]

Overall, it appears from this brief review that the sunk investment hypothesis can explain several of the main features of regulation that we observe in practice, without recourse to arguments based on distributional concerns, equity concerns, or "political" concerns.[62]

---

[60] Bonbright (1988, p. 384), in his list of attributes of a desirable tariff structure expresses this as the absence of "intercustomer burden".

[61] As in the example of the interstate rail track network in Australia which is subject to regulation despite the fact that (it claims) it is earning less than a normal return on its investments.

[62] Another interesting question is the extent to which this hypothesis could explain some of the key features of competition law. Despite attempts by economists to explain competition law on the basis of the notion of market power and deadweight loss, inconsistencies remain. Could it be that the primary rationale for, say, the control of mergers is to protect the sunk investment of users of the merging firms from the risk of expropriation resulting from the merger?

6 Conclusion

This paper has an ambitious objective – to call into question the conventional economic rationale for natural monopoly regulation. In my view, the hypothesis that the primary purpose of natural monopoly regulation is the minimization of deadweight loss 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 "market failure" approach to regulation?

Some economists have argued that regulation is best viewed as the outcome of political processes which are manipulated to serve private interests – such as the protection of the incumbent firm, or redistribution of wealth to consumers. The approach set out in this paper is complementary to these existing theories. Unlike that earlier literature, this paper starts from the hypothesis that there is an underlying welfare-maximizing rationale for regulation that is reflected in the patterns of regulation we observe over time and across countries. But what might that rationale for regulation be?

I have suggested that many features of natural monopoly regulation can be better understood as designed to promote and protect sunk complementary 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 only in the case of a monopoly is there a risk of expropriation of the value of these investments.

Viewing natural monopoly regulation in this way helps us to understand the patterns of regulation and the behavior of regulators that we see in practice. For example, it helps to explain opposition to price discrimination and Ramsey pricing, which are soundly based in conventional economic theory. It also helps to explain the heavy emphasis on price and service stability, and on incremental cost as a basis for pricing, even though these policies have little basis in conventional economic theory..

Economists have traditionally focused on the monopoly firm while assuming that all the relevant characteristics of the buyer side of the market can be expressed in the demand curve. This focus has neglected the important possibility that buyers of the monopoly services are not necessarily passive but often must take their own irreversible actions to extract the most value from the transaction. Taking these sunk investments into account allows us to go some way to bridging the still-broad gap between conventional economic theory and regulatory practice.

7 References

Andres, Luis, Jose Luis Guasch, and Stephane Straub, "Do Regulation and Institutional Design Matter for Infrastructure Sector Performance?", World Bank Policy Research Working Paper No. 4378, October 2007

Biggar, Darryl (1995), "A Model of Punitive Damages in Tort", *International Review of Law and Economics*, 15: 1-24.

Bonbright, James C., Albert L. Danielsen and David R. Kamerschen, (1988), *Principles of Public Utility Rates,* Washington DC, Public Utility Reports

Braeutigam, Ronald R.. (1989), "Optimal Policies for Natural Monopolies", chapter 23 in R. Schmalensee and R. D. Willig eds., *Handbook of Industrial Organisation*, Volume 2, page 1289.

Charles River Associates (2006), "Note on deregulation of rail infrastructure", 10 October 2006, available at www.pc.gov.au/__data/assets/file/0009/48861/subdd086.rtf

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

Coase, R. H., (1946), "The Marginal Cost Controversy", *Economica*, 13:169-182

Crocker, Keith and Scott Masten (1996), "Regulation and Administered Contracts Revisited: Lessons from Transaction-Cost Economics for Public Utility Regulation", 9: 5-39.

Decker, Chris (2007), "Bridging the Gap Between Economic Principle and Regulatory Convention: The Case of Ramsey Pricing", mimeo

Farrell, Joseph and Nancy T. Gallini (1988), "Second-Sourcing as a Commitment: Monopoly Incentives to Attract Competition", *The Quarterly Journal of Economics*, 103: 673-694

Faulhaber, Gerald (1975), "Cross-Subsidization: Pricing in Public Enterprises", *American Economic Review*, 65:966-977

Faulhaber, Gerald and William Baumol (1988), "Economists as Innovators: Practical Products of Theoretical Research", *Journal of Economic Literature*, 26: 577-600.

Goldberg, Victor (1976), "Regulation and administered contracts", *Bell Journal of Economics*, 7: 426-448

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

Holmes, Thomas (1990), "Consumer Investment in Product-Specific Capital: The Monopoly Case", *The Quarterly Journal of Economics*, 105: 789-801

Hotelling, Harold, (1938), "The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates", *Econometrica*, July 1938, 6, 242-269.

Joskow, Paul (1974), "Inflation and Environmental Concern: Structural Change in the Process of Public Utility Price Regulation", *Journal of Law and Economics*, 17: 291-327.

Joskow, Paul (1985), "Vertical Integration and Long-Term Contracts: The Case of Coal-Burning Electric Generating Plants", *Journal of Law, Economics and Organization*, 1: 33-80.

Joskow, Paul (1991), "The Role of Transactions Cost Economics in Antitrust and Public Utility Regulatory Policies", *Journal of Law, Economics and Organization*, 7, Special Issue, 1991, 53-83

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

Laffont, J.-J. and J. Tirole (1993), *A Theory of Incentives in Procurement and Regulation*, MIT Press

Laffont, J.-J. and J. Tirole (2000), *Competition in Telecommunications*, MIT Press

Leeson, Peter and Russell Sobel (2006), "Costly Price Discrimination", *Economics Letters*, 93

Leveque, Francois (2003), "Legal Constraints and Economic Principles", chapter 1 in *Transport Pricing of Electricity Networks*, Francois Leveque (ed), 2003, Kluwer Academic Publishers

Lynch, John, Thomas Buzas and Sanford Berg (1994), "Regulatory Measurement and Evaluation of Telephone Service Quality", *Management Science*, 40: 169-194

Milne, David, Esko Niskanen and Erik Verhoef, (2000), "Operationalisation of Marginal Cost Pricing within Urban Transport", VATT Research Reports 63, Government Institute for Economic Research (VATT)

Owen, Bruce and Ronald Braeutigam, (1978), *The Regulation Game: Strategic Use of the Administrative Process*, Cambridge

Peltzman, Sam, (1976), "Towards a More General Theory of Regulation", *Journal of Law and Economics*, 19: 211-240.

Priest, George (1993), "The Origins of Utility Regulation and the 'Theory of Regulation' Debate", *Journal of Law and Economics*, 36: 289-323

Spiller, Pablo T. and Mariano Tommasi (2005), "The Institutions of Regulation", in C. Menard and M. Shirley eds., *Handbook of New Institutional Economics*, 2005.

Stigler, George J., (1971), “The Theory of Economic Regulation”, *Bell Journal of Economics and Management Science*, 2:3-21

Train, Kenneth E. (1991), *Optimal Regulation: The Economic Theory of Natural Monopoly*, MIT Press

Vickery, William S. (1955), “Some Implications of Marginal Cost Pricing for Public Utilities”, *American Economic Review, Papers and Proceedings*, 45: 605-620

Viscusi, W. Kip, John M. Vernon, and Joseph E. Harrington, Jr. (1995), *Economics of Regulation and Antitrust*, 1995