The theory of access pricing and interconnection

Mark Armstrong

Nuffield College, Oxford

May 2001

The Theory of Access Pricing and Interconnection*

Mark Armstrong
Nuffield College
Oxford OX1 1NF, UK

May 2001

1 Introduction

This chapter discusses the interaction between competition and regulation in telecommunications markets, with the focus being on the central issue of access charges and network interconnection.[1] The main discussion is divided into three parts. In section 2 I discuss the problem of "one way" access pricing, which is where entrants need to purchase vital inputs from the incumbent, but not *vice versa*. In this case, because of the incumbent's monopoly position in the access market the firm behaves in many ways like a text-book monopolist, and without control it will usually set access charges too high. The desirability of regulatory intervention is therefore clear, and the section discusses how to set access charges in a variety of situations: when the incumbent's retail is chosen in advance (and perhaps in an *ad hoc* way); when the incumbent's retail prices and access charges are chosen simultaneously to maximize welfare (Ramsey pricing), and when the incumbent is free to choose its retail tariff.

In section 4 I discuss the "two way" access, or network interconnection, problem, which is where all firms in the market need to purchase vital inputs—namely, access to rival firms' subscribers—from each other. In this situation the danger might not be so much one of *foreclosure*, as in the one way case, but of *collusion* between networks. Can free negotiations between networks over their mutual access charges induce high prices for subscribers? The answer to this question is subtle, and depends in part on the kinds of tariffs that networks offer. Bridging these two sections is a discussion in section 3 of "competitive bottlenecks",

*The section on one way access pricing owes a great debt to joint work with John Vickers. I am also grateful to Eric Bond, Carli Coetzee, Wouter Dessein, Michael Doane, Joshua Gans, Martin Peitz, Patrick Rey, David Sappington, Vasiliki Skreta, Daniel Spulber, Jean Tirole, Tommaso Valletti, and Julian Wright for several helpful comments. I am responsible for all views and remaining errors.

[1]Another survey on the same topic can be found in Chapters 3, 4 and 5 of Laffont and Tirole (2000). Points of similarity and contrast between the two surveys will be discussed at various places in this chapter.

which is where networks operate in a competitive market *for* subscribers, and yet have a monopoly position for providing access *to* these subscribers. (This model is motivated by the mobile telephony market and the internet market.) Not surprisingly, this kind of market exhibits features both of competitive markets (no excess profits overall) and of monopoly markets (in the mobile sector, high profits are generated in the market for call termination, which are then used to attract subscribers).

Before starting the analysis, however, it is worth listing a few kinds of access services. There are numerous kinds of entry strategies a new telecommunications firm could follow—ranging from utilizing basic resale of an incumbent firm's own retail service, which involves little in the way of physical investment, to full-blown infrastructure-based entry—and each strategy requires its own kinds of access or interconnection services from incumbent firms. A very partial list of such services includes:

- Call termination: When a customer connected to firm A calls a customer of firm B, the former must (usually) pay the latter to deliver its call. Included within this is international call termination, where the network of a caller in one country must pay the relevant network in the destination country to deliver its calls (these payments being knows as "settlement rates" in this context).

- Call origination: A firm may have chosen not to construct a link directly to its customers, and so must rely on another firm to provide this link. An obvious example is long-distance competition, where a firm may choose to compete only on long-distance services, and therefore needs to pay an incumbent provider of local exchange services to originate (and terminate) its calls.

- Leasing network elements: A firm may lease some part of another firm's network—for instance, the local link into a subscriber's home (known as "local loop unbundling")—for some specified period. At one level this is just like call termination and origination except that the network is leased per month rather than per minute. However, contracts for local loop unbundling might allow the entrant to be more flexible in the technology it uses in combination with the network element—such as a different signalling technology—and so in practice this may be quite different.

- Roaming on mobile networks: For mobile networks, the equivalent of call termination and origination is "roaming", which is when a subscriber moves outside the area covered by her own network A, and still wishes to make and receive calls. In this case network A needs to arrange for a second network with adequate coverage to pick up and deliver the calls.

- Spectrum rights: A more esoteric "access" service is the need by mobile networks for suitable electromagnetic spectrum, a resource which is typically managed by government. The trend is for this essential facility to be auctioned off by government.

2 One-way Access Pricing

Many of the issues involving access pricing are best analyzed in a one way access framework, i.e. where the incumbent firm has a monopoly over important inputs needed by its rivals, but it itself needs nothing from other firms. (Outside the telecommunications sector, in markets such as gas and electricity, the one way paradigm is essentially the only one that is needed.) We discuss one way access pricing policy in a variety of situations. However, the analysis is entirely concerned with the case of vertical integration: the supplier of access services also operates in the retail markets supplied (potentially) by its rivals. The case of vertical separation is simpler, and has been analyzed elsewhere.[2] In addition, the focus is entirely on regulating access *charges*, and the important possibility that the incumbent will try to disadvantage rivals using various non-price instruments is ignored.[3]

Before we present the main analysis, some important preliminaries are discussed: how to achieve efficient entry when the incumbent's retail prices are out of line with its costs (section 2.1), and how an unregulated vertically integrated firm will behave towards its rivals (section 2.2). Although it is usual to treat the (regulated) access pricing problem and the (unregulated) foreclosure problem separately, it is one of the pedagogical aims of this chapter to show how many of the principles that underlie the policies that enable an unregulated firm to extract maximum profit from consumers and its rivals carry over to regulatory policies aimed at maximizing social welfare.

2.1 The Effect of Unbalanced Retail Tariffs

Incumbent telecommunications firms are often forced to offer retail tariffs that depart significantly from their underlying costs in several dimensions. (As we will see, the access pricing problem would be trivial if this were not the case.) There are two broad ways in which prices can depart from marginal costs. First, there is the problem caused by fixed and common costs: setting all prices equal to marginal cost will not allow the incumbent to break even. This Ramsey problem—which involves the calculation of the *optimal* departures of prices from costs—is discussed in section 2.5 below. Second, there is the problem that prices are determined in some, perhaps *ad hoc*, manner and may not reflect costs at all, and profits from one market are used to subsidize losses in others.[4] The focus in this section is on the second kind of "distortion".

[2]See for instance, section 5.2.1 in Armstrong, Cowan, and Vickers (1994) and section 2.2.5 in Laffont and Tirole (2000). Papers that discuss the disirability or otherwise of vertical integration include Vickers (1995), Weisman (1995), Sibley and Weisman (1998) and Lee and Hamilton (1999). (These latter papers however discuss unregulated downstream markets.)

[3]For further discussion of non-price discrimination against rivals, see section 4.5 in Laffont and Tirole (2000), Sibley and Weisman (1998), Economides (1998) and Mandy (2000).

[4]It is outside the scope of this paper to discuss *why* such cross-subsidies are so prevalent, or whether they are desirable. See Chapter 6 of Laffont and Tirole (2000) and Riordan (2002) for discussions of this, and for further references.

Examples of this practice in telecommunications include:

- A requirement to offer geographically uniform retail tariffs even though the costs of network provision vary across regions;
- A requirement to offer specific groups of subscribers subsidized tariffs;
- A requirement that call charges be used partly to cover non-traffic sensitive network costs, and that fixed charges do not cover the full costs of fixed network provision;
- One-off connection costs are not charged to subscribers as a one-off lump-sum, but much of the cost is collected periodically (e.g. quarterly) as fixed charges;
- Making a call often involves only a call set-up cost and the call cost is independent of duration, and yet call charges are typically levied on a per-minute or per-second basis.

Naturally such patterns of cross-subsidy lead to difficulties with laissez-faire entry, and there will tend to be "too much" entry into artificially profitable segments and "too little" in the loss-making markets. In addition there is the funding problem: if entry eliminates profits in the previously profitable markets, then the incumbent may be unable to continue to fund its operations in loss making markets. Since they have nothing to do with the presence of essential facilities *per se*, for maximum clarity I examine these issues in this section assuming that entrants do not need access to the incumbent's network to provide their services.[5]

In the next two sub-sections I discuss how efficient entry can be ensured using two different models for the competitive process. (These two models will be used throughout section 2 of the chapter.) Despite the apparent duplication, I use both models as they illustrate complementary aspects of competition in the market. The first model might be termed the "unit demands" (or "contestable") model, and it involves a one-for-one displacement of services from the incumbent to the entrant when entry occurs. The crucial aspect of this model is that, as long as prices do not exceed their gross utility, consumers have fixed and inelastic demands. This feature implies that there are no welfare losses caused by high retail prices (provided consumers are served). As a result, we will see that the incumbent will typically not act to distort the market, even if left unregulated, and this feature perhaps has unrealistic welfare implications. In addition, competition is "all or nothing" and either the incumbent *or* the entrant will serve a given subscriber group. This model is used mainly because it provides easy illustrative examples, and provides a simple intuitive way to discuss many of the main policy issues without getting distracted by demand elasticities, product differentiation, and so on.

The second model is termed the "competitive fringe" model, and involves the same service being offered by a group of entrants, where this service is differentiated from the incumbent's

[5]Thus, this section would apply naturally to postal service. See Crew and Kleindorfer (1998) for a related analysis.

own offering. Competition within the fringe means that prices there are driven down to cost and the fringe makes no profit. This model is more appropriate when entry is not all-or-nothing, so that the incumbent retains some market share even when entry occurs. Thus, this model works quite well when discussing competition in the long-distance or international call markets, when a large number of firms may enter the market and yet the incumbent retains a market presence. Because demand curves are downward sloping with this model, there are welfare losses associated with high prices, and this provides a rationale for regulatory intervention.[6]

2.1.1 Unit Demand Model

Consider a specific subscriber group that is offered a retail package by the incumbent, denoted M, which may be out of line with M's costs.[7] Suppose M incurs a total cost C per subscriber, and generates gross utility U per subscriber. The price for M's service is mandated to be P per subscriber (this price being determined by a process outside the model). A subscriber's net utility is therefore $U-P$. Suppose there is a potential entrant, denoted E, who can supply a rival package that costs c per subscriber and generates gross utility of u per subscriber.[8] Welfare per subscriber, as measured by the sum of consumer utility and profits, is equal to $u-c$ if E serves subscribers and $U-C$ if M retains the market. Therefore, successful entry is socially desirable if and only if

$$C \geq c + [U - u] \ . \tag{1}$$

Given M's price P, the entrant can attract subscribers provided its own price p is such that $u-p \geq U-P$. Therefore, entry will occur whenever the maximum price that can be charged by E, which is $p = P - [U - u]$, covers its costs, i.e. when

$$P \geq c + [U - u] \ . \tag{2}$$

[6]Note that neither of these models allow the entrant(s) to have any effective market power, and this assumption is maintained for most of the chapter. (In the section on two way interconnection, there are several cases where all firms have market power.) If entrants do have market power then access charges should be chosen with the additional aim of controlling the retail prices of entrants. This would typically lead to access charges being set lower than otherwise, following the same procedure as the familiar Pigouvian output subsidy to control market power. Allowing for this will make an already complex discussion more opaque. However, see section 3.3.1 of Laffont and Tirole (2000) for a discussion of the implications for policy when entrants have a degree of market power.

[7]This discussion, as well as those in sections 2.3.2 and 2.4.1, is taken from Armstrong (2001). The geographical example is used because it is the simplest way to make the main points. However, the same broad principles apply to other ways of partitioning subscribers, such as into high-usage and low-usage groups. The complication of this kind of alternative framework is that firms will most likely not be able directly to observe the group to which subscribers belong, and so will have to *screen* subscribers—see section 3.2.3.3 of Laffont and Tirole (2000) for analysis along these lines.

[8]This utility could be higher than that supplied by M (if E uses a newer and superior technology for instance), or it could be lower (if subscribers incur switching costs when they move to the entrant).

Whenever $P \neq C$, therefore, private and social incentives for entry differ.

There are two kinds of market failure, depending on whether the sector is profitable or loss-making for the incumbent. Suppose first that the market segment is required to be profitable, so that $P > C$. Then whenever

$$P \geq c + [U - u] \geq C$$

entry occurs when it is socially undesirable. In this case entry can profitably take place even when the entrant has higher costs and/or lower quality than the incumbent. (In this sense there is "too much" entry.) Alternatively, if $P < C$ then when

$$P \leq c + [U - u] \leq C$$

it is socially desirable for entry to take place, and yet it is not privately profitable. In this case entry does not occur even though the entrant has lower costs and/or a higher quality of service. In sum, whenever the incumbent's prices are required to depart from its costs of serving subscribers, there is a danger of undesirable entry into profitable markets and of too little efficient entry into loss-making markets.

In theory it is a straightforward matter to correct this divergence between the private and social incentives for entry. For these subscribers the incumbent is implicitly paying an "output tax" of

$$t = P - C \tag{3}$$

per subscriber—which is positive or negative depending on whether the segment is profitable—and efficient entry is ensured provided that the entrant is *also* required to pay this tax.[9] Notice that this tax is equal to the incumbent's lost profit—or "opportunity cost"—when it loses a subscriber to its rival.

While it may seem a little abstract to use these kinds of output taxes to correct for allocative inefficiencies in the incumbent's tariff, these can sometimes be implemented in a simple and non-discriminatory way via a well designed "social obligations" or "universal service" fund of the kind sometimes proposed for the telecommunications industry. This procedure can be illustrated by means of a simple example, summarized in Table 1 below. (I return to variants of this example later in the chapter.)

[9] With this tax an entrant will find it profitable to take over a subscriber provided that

$$u - c - t \geq U - P \,,$$

i.e. whenever $u - c \geq U - C$. Therefore, entry takes place if and only if it is desirable.

	URBAN	**RURAL**
number of subscribers	20m	10m
M's cost per subscriber	\$50	\$200
M's price per subscriber	\$100	\$100
M's overall profit for each type	\$1bn profit	\$1bn loss
Any firm's contribution to fund	\$50	-\$100

Table 1: Giving correct entry incentives via a universal service fund

Here, the incumbent offers a retail service to two groups of subscribers, a high cost rural group and a low cost urban group. Regulatory policy, in the form of universal service obligations, demands that the incumbent offers service to both groups at the same price, \$100, and (without entry) the firm makes a profit from urban subscribers that just covers the loss from rural subscribers.

As discussed above, a *laissez-faire* entry policy will most likely lead to (i) inefficient entry into the artificially profitable urban market, (ii) too little efficient entry into the loss-making rural sector, and (iii) funding difficulties for the incumbent in the event of cream-skimming urban entry. To counter these problems, suppose the regulator sets up a fund containing \$1bn to finance the rural sector. The source of this fund is the profits generated in the urban sector, and either firm—the entrant or the incumbent—must pay an amount \$50 (M's profit margin in this sector) into this fund for each urban subscriber they serve. In return, any firm that serves a rural subscriber receives a subsidy from the fund equal to \$100 (M's per-subscriber loss in that sector) for each subscriber served. By construction, provided the number of subscribers in the two groups does not change with entry, such a fund is self-financing, and widespread entry does not undermine the ability of the incumbent to supply the loss making market. More important from an economic efficiency point of view, though, is the feature that the contribution scheme ensures that in each sector the entrant has to pay the output tax/subsidy (3) which gives it the correct entry incentives.[10]

2.1.2 Competitive Fringe Model

The second model of competition shows how this policy applies in a different context.[11] As discussed above, in order to introduce product differentiation, and to sidestep the issue of the market power of entrants, I assume there is a competitive fringe of price-taking entrants, still denoted E, all of whom offer the same service (which is differentiated from that supplied by M). The impact of modelling competition as a competitive fringe is that we may assume the rivals' price is always equal to their perceived marginal cost, and the fringe makes no profits.

[10] Clearly, in implementation great care must be taken to ensure that entrants cannot "bypass" this output tax, for instance by providing a similar but not identical service.

[11] This model of competition is adapted from Laffont and Tirole (1994) and Armstrong, Doyle, and Vickers (1996).

Let P and p be M's price and E's price for the respective services. Let $V(P,p)$ be consumer surplus with the two prices. This satisfies the envelope conditions $V_P(P,p) = -X(P,p)$ and $V_p(P,p) = -x(P,p)$, where X and x are, respectively, the demand functions for the incumbent's and the fringe's services.[12] Assume that the two products are substitutes, so that $X_p \equiv x_P \geq 0$. As before, the incumbent has marginal cost C and the fringe has marginal cost c. In order to achieve the correct amount of entry, the regulator levies a per unit output tax t on the fringe's service. Then competition implies that the fringe's equilibrium price is $p = c + t$. Keeping the incumbent's price fixed at P, the regulator aims to maximize total surplus (including tax revenue) as given by

$$W = \underbrace{V(P, c+t)}_{\text{consumer surplus}} + \underbrace{tx(P, c+t)}_{\text{tax revenue}} + \underbrace{(P-C)X(P, c+t)}_{M\text{'s profits}} . \tag{4}$$

Maximizing this with respect to t implies that the optimal fringe price and output tax is given by

$$p = c + \sigma_d(P - C) \; ; \; t = \sigma_d(P - C) \; , \tag{5}$$

where

$$\sigma_d = \frac{X_p}{-x_p} \geq 0 \tag{6}$$

is a measure of the substitutability of the two products: it measures how much the demand for M's service decreases when E supplies one further unit of its own service.[13] In particular, if the market is profitable, i.e. when $P > C$, then it is optimal to raise the rivals' price above cost as well, i.e. to set $t > 0$. The reason for this is that profits are socially valuable, and given $P > C$ it is optimal to stimulate demand for M's service, something that is achieved by driving up the rival price (given that services are substitutes).

Analogously to (3) above, t is M's lost profit caused by the *marginal* unit of fringe supply. This lost profit is the product of two terms: the marginal profit $(P - C)$ per unit of final product sales by the incumbent, and σ_d which is the change in the incumbent's sales caused by increasing fringe output by one unit. Therefore, (5) indeed gives the loss in the incumbent's profit caused by E's marginal unit of supply. With perfect substitutes we have one-for-one displacement of rival for incumbent services, i.e. $\sigma_d = 1$, in which case the rule reduces to the simple formula (3). If the products are not close substitutes, so that σ_d is close to zero, then this optimal tax should also be close to zero, and a laissez-faire entry policy is almost ideal. This is because policy towards the fringe market has little impact on the welfare generated in the incumbent's market, and therefore there is no incentive to set a price in the fringe market different from cost.

[12] Subscripts are used to denote partial derivatives. I assume that consumers have no "income effects", so that Roy's identity reduces to these envelope conditions.

[13] Given P, if E supplies a further unit of service its price must decrease by $1/x_p$, which in turn causes M's demand to fall by X_p/x_p. Note that in general σ_d is a function of the tax t, and so expression (5) does not give an explicit formula for the tax. However, provided the demand functions are reasonably close to being linear—so that σ_d is close to constant—this issue is not important.

2.1.3 Discussion: The Theory of the Second-Best

The rules (3) and (5) are instances of the "general theory of the second best", which states that if one service is not offered at the first-best marginal cost price ($P \neq C$), then the optimal price in a related market also departs from marginal cost ($p \neq c$).[14] Therefore, taxes of the form (3) or (5) are examples of what might be termed the "second-best output tax". Indeed the optimal policy in this context states that the correct departure from marginal costs should be given by the incumbent's lost profit—or opportunity cost—due to entry into its market. We can sum up this discussion by the formula:

$$\begin{aligned} &\textit{Second-best output tax required to implement efficient entry} \\ &\qquad = \textit{incumbent's lost profit in retail market .} \end{aligned} \tag{7}$$

From an economic efficiency point of view it makes little difference whether the proceeds from this tax are paid directly to M, to the government, or into an industry fund. However, if the incumbent has historically been using the proceeds from a profitable activity to finance loss-making sectors then if the entrant pays the tax to the incumbent, the incumbent will not face funding problems if entry takes place. However, perhaps a more transparent mechanism would be for a "universal service" fund to be used to finance loss-making services, as illustrated by Table 1 above. Notice that if the charge *is* paid to the incumbent firm then the firm is indifferent to whether it retains or loses market share to a rival. This has the political benefit that, with such a tax regime, the incumbent has no incentive to lobby against entry.[15] However, from an efficiency point of view this is irrelevant, and it is merely a happy feature of the regime that this tax which gives correct entry incentives also recompenses the incumbent for its loss of market share.

2.2 The Problem of Foreclosure in an Unregulated Market

Does a vertically integrated firm with a monopoly over vital inputs have an incentive to distort competition in related markets? The so-called foreclosure (or essential facilities) doctrine says *Yes*, whereas the so-called Chicago critique claims *No*.[16] It is not the aim of this section to probe this general issue in any depth. Rather, I wish to demonstrate that the main principles that govern the unregulated firm's behavior toward access pricing—or

[14] See Lipsey and Lancaster (1956).

[15] Or at least this is true in the short run; in the longer term the incumbent might believe that entrants will enjoy future benefits (at the incumbent's expense) once they have a firm presence in the market. In addition, the fact that the incumbent's profits do not depend on the extent of entry implies that it will not have any (short-run) incentive to foreclose entry through non-price means.

[16] For a discussion of the topic of market foreclosure, see Rey and Tirole (2006) and Vickers (1996). In this paper I look at the incentive for an already vertically integrated firm to distort the downstream market, whereas most discussions of foreclosure concentrate on the case of vertical separation and the inefficiencies that result from this. (Vertical integration is then often a *solution* to the problem caused by these inefficiencies.)

wholesale pricing as it is usually termed in the unregulated case—are closely related to the regulator's incentives for choosing the access charge. In particular, the number of instruments available to the firm/regulator affects whether or not productive efficiency is desirable.

Consider this simple framework: there is one vertically integrated firm M and a rival firm (or group of firms) E which may need access to M's "upstream" input to be able to compete with M at the "downstream" retail level. Firm M has constant marginal cost C_1 for providing its end-to-end retail product and C_2 for providing its upstream access product to E. Firm M chooses the per-unit charge a for its access service.[17] As in section 2.1 above, we divide the analysis into two parts, depending on the nature of the downstream competitive interaction. In addition, we consider cases where the entrant can "bypass" M's upstream input—i.e. manage to supply its retail service without the use of this input from M—either by finding an alternative supplier for this service or by providing the service itself.

2.2.1 Unit Demand Model

As in section 2.1.1, suppose that M's retail service generates gross utility of U per unit, and that E's service (when it uses M's access service) generates utility u. Suppose that it costs E an additional amount c to transform a unit of M's access service into a unit of its retail service.

No possibilities for bypass: Suppose first that E definitely requires M's access service to provide its retail service. Therefore, as in (1) above, it is socially desirable for E to enter whenever $u - [c + C_2] \geq U - C_1$, i.e. when

$$C_1 - C_2 \geq c + [U - u] \ . \tag{8}$$

Is it in M's selfish interest to distort the market, and to foreclose the rival even if it is efficient for the entrant to serve the market? In this special framework the answer is *No*. First notice, as in (2), that if M chooses the pair of prices P and a, where the former is its retail price and the latter is the access charge, then the entrant will choose to enter if and only if the margin $P - a$ satisfies

$$P - a \geq c + [U - u] \ . \tag{9}$$

(If M chooses the retail price P then to attract subscribers E must offer a price p such that $u - p \geq U - P$. It will choose to enter if a price that covers its total cost $c + a$ satisfies this condition.) If M does shut out E, for instance by setting a very high access charge a, then its maximum profit is obtained by charging the maximum acceptable retail price, which is $P = U$, and this generates profits of $U - C_1$ per unit. On the other hand, if it allows E to provide the retail product, for instance by setting a very high retail price P, then it can set

[17] More generally, M will want to choose more complicated access tariffs, including franchise fees and so on, to extract any available profits from E. However, in our simple models the rivals necessarily make no profit and so there is no need to use additional instruments for this purpose.

an access charge up to $a = u - c$ (which just allows E to break even at its maximum possible price $p = u$), which gives M a profit of $u - c - C_2$ per unit. Therefore, by comparing these profits with the condition for efficiency in (8) we see that M will allow entry if and only if this is efficient. The incumbent manages to obtain the full surplus, and subscribers and E are left with nothing.

Thus the firm has no incentive to distort competition downstream. Essentially this is an instance of the "Chicago critique" of foreclosure, which in simple terms states that vertical integration has no anti-competitive effects. In this case, because of its bargaining power, the unit-demand nature of subscriber demand and the assumption that M has full information about the entrant's technology, M can obtain E's retail service at marginal cost, and so will use this segment whenever its own (quality adjusted) cost is higher.[18] However, as is well known, this result is non-robust in a number of ways—see the next sub-section for one example. One way in which this insight does extend, though, is to the potential for bypass on the part of the entrant.

Bypass: Suppose next that the entrant can provide its own access service, thus bypassing M's network altogether.[19] Suppose that when it provides its own access it incurs total costs $\hat{C}_1$ per unit for its end-to-end service. This new network gives subscribers a gross utility $\hat{u}$. (Utility $\hat{u}$ may differ from u if using the incumbent's network degrades or enhances the entrant's service compared to its stand-alone service.) Therefore, the entrant can charge a price $p = P + [\hat{u} - U]$ and attract subscribers, without any need for access to M's upstream service. Total surplus per subscriber with this mode of entry is just $\hat{u} - \hat{C}_1$. In sum, social welfare per unit with the three possible entry strategies is

$$W = \begin{cases} \hat{u} - \hat{C}_1 & \text{with stand-alone entry} \\ u - c - C_2 & \text{with entry using } M\text{'s access service} \\ U - C_1 & \text{with no entry .} \end{cases} \tag{10}$$

Given the incumbent's pricing policy (P, a), which of the three options will the entrant follow? If it decides to use M's access service it can charge up to $p = P + [u - U]$ and still attract customers, in which case it makes a profit per unit of $P + [u - U] - [a + c]$. On the other hand, if E bypasses M's network it can make profit of $P + [\hat{u} - U] - \hat{C}_1$. Therefore, *given* that E enters we deduce that it will use M's network provided that

$$a \leq [u - \hat{u}] + [\hat{C}_1 - c] \ , \tag{11}$$

regardless of M's retail price. (The price P affects the profitability of entry itself, but not

[18] Exactly the same efficiency effect is at work when a text-book monopolist faces consumers with unit demands (with known reservation prices). The monopolist will simply charge the maximum price consumers will pay, which causes no deadweight losses due to the shape of the demand functions.

[19] I ignore the possibility that the entrant should build a network and the incumbent should provide retail services over this new network, i.e. that the entrant should provide "access" to the incumbent.

the relative profitability of the two modes of entry.) Therefore the maximum profit per unit that M can make when it succeeds in supplying access to E is $[u - \hat{u}] + [\hat{C}_1 - c] - C_2$.

On the other hand, if it does not supply access to E what is M's maximum profit? Given that E can enter at cost $\hat{C}_1$ without the need for access, the maximum retail price that M can charge and yet not induce stand-alone entry is $P = \hat{C}_1 + [U - \hat{u}]$, which gives it profits of $[U - \hat{u}] + [\hat{C}_1 - C_1]$. In sum, M's maximum profit per unit under the three entry regimes are

$$\Pi = \begin{cases} 0 & \text{with stand-alone entry} \\ [u - \hat{u}] + [\hat{C}_1 - c] - C_2 & \text{with entry using } M\text{'s access service} \\ [U - \hat{u}] + [\hat{C}_1 - C_1] & \text{with no entry .} \end{cases} \tag{12}$$

Therefore, comparing (12) with (10) we see that the incumbent's incentives are precisely in line with overall welfare. In sum, in this model the potential for bypass does not affect the incumbent's incentive to allow the most efficient method of production. However, bypass *does* affect the profits that M can obtain: comparing (12) with (10) we see that M cannot appropriate all the surplus in (10), and $\hat{u} - \hat{C}_1$ of this surplus leaks out (either to E or to subscribers, depending upon whether entry takes place or not).

In the next section, with the alternative model of competition, we will see that the possibility of bypass *does* cause the incumbent to distort competition, at least when the incumbent can use only a limited set of instruments.

2.2.2 Competitive Fringe Model

Suppose next that E is a competitive fringe as in section 2.1.2 above. Suppose that with the access charge a the fringe has the (minimum) constant marginal cost $\psi(a)$ for producing a unit of its final service, including the payment of a per unit of access to M. If the fringe cannot bypass M's access service, so that exactly one unit of access is needed for each unit of its final product, then $\psi(a) = a + c$, say, where c is E's cost of converting the input into its retail product. Otherwise, though, E can substitute away from the access product, in which case $\psi(a)$ is concave in a.

Again we write P for M's retail price and p for E's retail price, and the demand for the two services is $X(P, p)$ and $x(P, p)$ respectively. Note that $\psi'(a)$ is, by Shephard's Lemma, E's demand for access per unit of its retail service, and therefore the total demand for M's access service is $\psi'(a)x$. Suppose for now that M can somehow levy a per-unit charge t on the output of the fringe. Since E is a competitive fringe its equilibrium retail price is equal to total perceived marginal costs: $p \equiv t + \psi(a)$.

Putting all of this together implies that M's total profits, Π, are comprised of three parts:

$$\Pi = \underbrace{(P - C_1)X(P, t + \psi(a))}_{M\text{'s profits from retail}} + \underbrace{(a - C_2)\psi'(a)x(P, t + \psi(a))}_{M\text{'s profits from access}} + \underbrace{tx(P, t + \psi(a))}_{\text{profits from output levy}} . \tag{13}$$

Alternatively, since $p = t + \psi(a)$ we can think of M choosing p rather than t, in which case this profit becomes

$$\Pi = (P - C_1)X(P, p) + \{p - \psi(a) + (a - C_2)\psi'(a)\}\, x(P, p)\ . \tag{14}$$

Clearly, whatever choice is made for the two retail prices, profit is maximized by choosing a to maximize the term $\{\cdot\}$ above, which has the first-order condition $[a - C_2]\psi''(a) = 0$. In particular, whenever there are possibilities for the fringe to substitute away from M's access product, so that $\psi'' \neq 0$, then $a = C_2$ at the optimum. Therefore, marginal cost pricing of access is the profit-maximizing strategy for a vertically-integrated firm when it can levy an output tax on its rivals. (The two retail prices P and p are then chosen by M to maximize its profits in (14).) In this framework, then, M has no incentive to distort rival supply in the sense of causing productive inefficiency.[20]

An often more realistic case is where M cannot impose an output tax on the fringe, so that $t = 0$ in (13).[21] In this case the firm chooses P and a to maximize its profits

$$\Pi = \underbrace{\Pi^R(P, a)}_{\text{profits from retail}} + \underbrace{(a - C_2)z(P, a)}_{\text{profits from access}} \tag{15}$$

where $\Pi^R(P, a) \equiv (P - C_1)X(P, \psi(a))$ is M's profit in the retail sector and $z(P, a) \equiv \psi'(a)x(P, \psi(a))$ is the fringe demand for access. (Throughout the chapter, z is used to denote the demand for access.) The solution to this problem has first-order conditions

$$\Pi^R_P + (a - C_2)z_P = 0\ ; \tag{16}$$

$$\Pi^R_a + z + (a - C_2)z_a = 0\ .$$

Notice that $a = C_2$ cannot satisfy these conditions, and the firm will choose to set $a > C_2$.[22] Therefore, when it has only the access charge as its instrument it is optimal for M to cause a degree of productive inefficiency within the fringe in order to drive up the rival retail price. In contrast to the simple unit demands model in section 2.2.1, the vertically integrated firm here *does* have an incentive to distort the production decisions of its rivals.

[20] Obviously, however, since both P and p will be set too high from a welfare perspective, the retail market is distorted in the sense of there being allocative inefficiency.

[21] It is unreasonable to suppose that the entrant can be forced to pay the incumbent an output charge if it chooses to supply its own stand-alone service. However, a potentially reasonable contract offered by the incumbent might take the following form: I offer you access at a per unit tied to your acceptance of paying me a per-unit output charge t. For simplicity, though, we assume here that M cannot levy any form of output charge on the fringe, perhaps because of the possible difficulty in observing or verifying the output of the fringe.

[22] It is straightforward to show that $z_a < 0$ and, provided $P > C_1$, that $\Pi^R_a > 0$. Therefore, suppose that $a \leq C_2$ solves these first-order conditions. For M to make any profits we must have $P > C_1$ and so $\Pi^R_a > 0$. The second of the first-order conditions is therefore a contradiction, and we deduce that $a > C_2$.

2.2.3 Discussion: Instruments and Objectives

In general the vertically integrated firm has three objectives when trying to maximize its profits: it wishes (i) to ensure there is productive efficiency, i.e. that industry costs are minimized; (ii) to maximize total industry profit given these minimized costs, and (iii) to extract all industry profit for itself. Consider the competitive fringe model of section 2.2.2. First, note that (iii) is automatically satisfied since competition within the fringe drives profits there to zero. When the incumbent can levy an output tax on the fringe and so has three instruments—P, a and t— we have seen that the remaining two objectives (i) and (ii) can be achieved. This is not surprising: because of product differentiation, objective (ii) requires two instruments, namely the control of the two retail prices; productive efficiency is obtained by the use of a further instrument, the access charge. By contrast, with only two instruments—for instance, if the output levy is not available— then (i) and (ii) cannot both be satisfied, except in the special case where bypass is not possible. (Clearly, when bypass is impossible then (i) is automatically satisfied.) In general, then, a compromise is needed: the access charge is required to control both productive efficiency (which suggests that the charge should be close to marginal cost) and the fringe's retail price (a high access charge leads to high retail prices). The optimal access charge, therefore, lies above cost and there is a degree of productive inefficiency.

The same effects will be seen to be at work in the following sections on regulated markets. The difference is that the regulator, not the firm, is controlling the access and other charges, and wishes to pursue two objectives: (i) to ensure there is productive efficiency, as in the unregulated case, and (ii) to maximize total welfare subject to these minimized costs. Again, if the regulator has enough instruments then both of these objectives can be achieved. In other cases, though, the access charge is required to perform too many tasks at once, and productive efficiency will most likely suffer.[23]

2.3 Fixed Retail Prices With No Bypass: The ECPR

In this section and the next, we focus on the sub-problem of how best to determine access charges for a *given* choice of the incumbent's retail tariff. (This retail tariff is assumed to be chosen by some regulatory process outside the model.) From an economic efficiency point of view, it is clearly a superior policy to consider the incumbent's retail prices and access charges simultaneously, since that allows for the various tradeoffs between consumer welfare and productive efficiency to be considered correctly—see section 2.5 for this analysis. However, it seems worthwhile to analyze this case of fixed and perhaps inefficient retail tariffs, since it is often the case that retail tariffs are not set according to strict Ramsey principles (at least as conventionally applied), and various political or historical considerations often have an crucial impact on the choice of the retail tariff.

[23] See section 3.3 in Laffont and Tirole (2000) for a more detailed discussion along these lines.

2.3.1 Different Representations of the ECPR

Probably the single most contentious issue in the theory and practice of access pricing is the status of the so-called *efficient component pricing rule*, or ECPR, which states that the access charge should, under certain circumstances to be determined, satisfy the formula:[24]

$$\begin{aligned} \textit{access charge} \;&=\; \textit{cost of providing access} \\ &\quad + \textit{incumbent's lost profit in retail markets caused by providing access .} \end{aligned} \tag{17}$$

What is meant by this becomes clearer if we use the notation introduced in section 2.2. Using the previous notation, the incumbent's access charge is a, its retail price is P, its marginal cost of supplying its downstream service is C_1, and its marginal cost of providing access to its rivals is C_2. Then the expression (17) can be expressed more formally as

$$a = C_2 + \underbrace{\sigma(P - C_1)}_{M\text{'s lost retail profit}} \quad . \tag{18}$$

Here the parameter σ measures how many units of M's retail service are lost by supplying a unit of access to its rivals.[25] In fact we have already seen this rule used implicitly in section 2.1. In that section there was no vertical element present, and so there was no direct cost of providing access, i.e. $C_2 = 0$. In the unit demands framework, one unit of "market access" enabled one unit of rival service to be supplied, which in turn displaced one-for-one a unit of the incumbent's retail service. Therefore, $\sigma = 1$ in this case and the ECPR proposal reduces to (3). Similarly, in the competitive fringe model, one unit of market access again enabled a unit of fringe service to be supplied, but this caused only σ_d units of incumbent service to be given up, which then gives (5).

From this perspective, using the identity (7) we can re-write (17) to give the perhaps less "incumbent friendly" formula:

$$\begin{aligned} \textit{access charge} \;&=\; \textit{cost of providing access} \\ &\quad + \textit{second-best output tax on entrants .} \end{aligned} \tag{19}$$

Although it is largely an issue of semantics, this way of expressing the formula better reflects the fact that departures from cost-based access pricing are the result of second-best corrections to account for the incumbent's distorted retail tariff, rather than the result of the need to compensate the incumbent for lost profit. The equivalence (19) explains why many of the implications of the analysis in this chapter coincide with those generated from

[24] This rule appears to have been proposed first in the pioneering analysis in Willig (1979)—for instance, see his expression (72). See also Baumol (1983), Baumol and Sidak (1994a), Baumol and Sidak (1994b), Baumol, Ordover, and Willig (1997) and Sidak and Spulber (1997a) for further discussions concerning the desirability of this rule.

[25] The parameter σ was termed the *displacement ratio* by Armstrong, Doyle, and Vickers (1996).

the very different approach of Sidak and Spulber (1997a). I discuss policy from the point of view of static economic efficiency, whereas Sidak and Spulber emphasize more the need to compensate the incumbent adequately for past investments when its network is opened up to rivals.[26]

Another formula that often goes under the name of the ECPR, but what is perhaps better termed the "margin rule", states that

$$\begin{aligned} \textit{access charge} &= \textit{incumbent's retail price} \\ &\quad - \textit{incumbent's cost in the retail activity} \ . \end{aligned} \tag{20}$$

Re-arranging, this rule states that the "margin" available to the rivals, i.e. the incumbent's retail price minus its access charge, is equal to the incumbent's cost in the competitive activity. Using the above notation, if C_2 is M's marginal cost of supplying access to itself (as well as to its rivals), then $[C_1 - C_2]$ is M's marginal cost in the competitive activity. Therefore, this rule is more formally expressed as

$$a = C_2 + [P - C_1] \ . \tag{21}$$

Clearly this formula coincides with (18) in the case where one unit of access supplied to the rivals causes a unit reduction in the demand for the incumbent's retail service, i.e. when $\sigma = 1$. Although it is again merely a question of semantics which of the rules (17) or (20) is termed the "ECPR", it is very important when it comes to discussing the applicability of the ECPR to distinguish between these two rules. Since we will see that the former is much more generally valid, we will use the term "ECPR" for that rule, and the term "margin rule" for the latter.[27]

We have already seen in section 2.2 that the level of the optimal access charge is going to depend crucially upon whether the regulator can impose an output tax on entrants. We will see that, suitably interpreted, the ECPR as expressed in (17) or (19) is the correct rule when additional instruments, such as output taxes on entrants, are not employed. When output taxes are used, however, then the ECPR is not appropriate, and pricing access at cost is the better policy (just as in the unregulated case in section 2.2).

[26] However, this focus on avoiding "deregulatory takings" could itself be viewed as promoting *dynamic* efficiency—see pages 214–216 in Sidak and Spulber (1997a).

[27] The discussion in section 3.2.5 of Laffont and Tirole (2000) is confusing from this perspective in that the margin rule and the ECPR are taken to be the same. For instance, the justification the margin rule on page 119 is given as (using the notation of this paper): "suppose that an entrant diverts one minute of long-distance phone call from the incumbent. This stealing of business costs the incumbent an amount equal to the markup on his long-distance offering, $P - C_1$, plus the marginal cost of giving access, C_2, to the entrant." This discussion assumes, though, that to steal one unit of business from the incumbent requires the rival to supply exactly one further unit of its own supply, whereas in general with imperfect substitutes the rival will need to supply more or less than the incumbent's lost demand.

A final point is that in symmetric situations with product differentiation, the Ramsey optimum involves the margin rule (but never the ECPR in our terminology) being satisfied—see section 2.5.1 below.

2.3.2 Unit Demands and the Margin Rule

Consider first the simple unit demands framework of section 2.2.1 above. The incumbent has constant marginal cost C_1 for its retail service and constant marginal cost C_2 for its access service. Its retail service generates gross utility U to each of its subscribers, and it is required by regulation to charge the retail price P (this price being determined by a process outside the model). As assumed throughout this section, bypass of the incumbent's network by rivals is not possible, and so (by a suitable choice of units) the entrant needs precisely one unit of access for each unit of its retail service.

There will be entry with the access charge a provided condition (9) holds. Total welfare is higher if the entrant supplies the market if and only if condition (8) holds. Therefore, entry incentives coincide with overall social welfare provided the access charge is determined by the "margin rule" (21). As explained above, this formula is the natural interpretation of the ECPR formula in (17) and (19) in this special context. Note that in this scenario where there is no possibility for the entrant to substitute away from M's access service, there is no productive inefficiency caused by setting the access charge above cost, and hence there is no need for additional instruments such as output taxes to achieve the correct outcome.

It is also worth noting that in this model there is actually no need, from an efficiency point of view, to control the incumbent's access charge (provided the incumbent knows the potential entrant's cost c and service quality u). For if M can choose its own access charge, for fixed P, it will choose to allow entry if and only if the maximum profits from selling access to the entrant are higher than the profits obtained when it sells the retail product itself. But the maximum profits available from selling access are obtained by setting a to be the highest value that satisfies (9), which gives it profits of $P - [c + C_2] - [U - u]$ per unit. Comparing this with the profit obtained by preventing access, which is just $P - C_1$, we see that the incumbent will allow entry to occur if and only if (8) holds, i.e. if entry is more efficient. Note that the level of its regulated retail price does *not* bias the incumbent at all in its dealings with its rival since P affects its profits from selling access and from foreclosing entry in exactly the same way. However, it is important for this result that M knows all relevant information about E's service, something that it unlikely to be the case in practice (since entry has not taken place at the time the incumbent has to choose its access charging policy). The advantage of the rule (21) is that it ensures efficient entry for all kinds of entrant, and does not require that the parameters c and u be known.

Finally, we can illustrate this margin rule in a simple extension of the example used in Table 1.[28] Again, there are two groups of subscribers, rural and urban, and the relevant data is summarized in Table 2. Here there are two components to providing a final service: the network element and the retail element. The incumbent is assumed to incur the same retail cost for each subscriber group, but its network cost differs across the two groups. (The incumbent's total end-to-end cost for providing service, denoted C_1 in the preceding analysis, is therefore the sum of these two terms.) Here the entrant cannot build its own network

[28] Baumol (1999) provides a similar argument to that in the following discussion.

and can compete only in the retail segment. As in section 2.1.1, the incumbent is forced by policy to charge the same amount to both groups, despite the underlying cost differences.

	URBAN	**RURAL**
number of subscribers	20m	10m
M's cost per subscriber, of which	\$50	\$200
retail cost is	\$20	\$20
network cost is	\$30	\$180
M's retail price for service	\$100	\$100
M's network access charge	\$80	\$80

Table 2: The optimality of the margin rule with no bypass

The margin rule (21) implies that the correct network access charge is \$80 per subscriber, regardless of the type of subscriber.[29] This access charge implies that entry will be profitable only if the entrant has retail costs lower than the incumbent's (or provides a superior service). This policy contrasts with a "cost-based" access charging policy, which would require charging for urban access at \$30 and charging for rural access at \$180, a policy that leads to precisely the same problems as indicated in section 2.1.1. For instance, with a network access charge of \$30 for urban services, an entrant could have a retail cost of as high as \$70 (as compared to the incumbent's retail cost of \$20) and still find entry profitable.

2.3.3 Competitive Fringe Model

Next, consider the model of the industry as described in section 2.2.2 above, simplified so that there is no possibility of bypass. By making a suitable choice of units, we can assume that the fringe needs exactly one unit of access for each unit of final product it supplies. Therefore, the perceived marginal cost of the fringe is $a+c$, where c is E's cost of converting access into its retail product, and so competition within the fringe implies that E's equilibrium retail price is also $a+c$. As in (4) above, welfare with the access charge a is

$$W = \underbrace{V(P, c+a)}_{\text{consumer surplus}} + \underbrace{(a - C_2)x(P, c+a)}_{M\text{'s profits from access}} + \underbrace{(P - C_1)X(P, c+a)}_{M\text{'s profits from retail}} . \tag{22}$$

As in (5), maximizing this with respect to a gives the following expression for the optimal access charge:

$$a = C_2 + \sigma_d(P - C_1) , \tag{23}$$

[29] The fact that both the retail charge and the retail cost are uniform in this example implies that the margin rule access charge will also be uniform. In fact the access charge, \$80 applied uniformly, is just the geographically averaged network cost in this example, a feature that follows from the assumption that the incumbent is regulated to make zero profits overall.

where σ_d is given in (6) above. (By inspection, the regulator's problems in (4) and (22) are identical, where the "tax" t in the former expression represents the markup $a - C_2$ and the tax revenue in the former plays the same role as the incumbent's profits from access in the latter.) Clearly this is the appropriate version of (17), and so the ECPR is again the correct rule in this framework. Following the discussion in section 2.3.1 above, however, we see that the margin rule (20)—(21) is *not* valid in this framework.

In the special case where the two retail services are approximately independent in terms of consumer demand, so that $\sigma_d \approx 0$, there is no opportunity cost to the incumbent in providing access. In particular, the ECPR rule (23) states that the access charge should then involve no mark-up over the cost of providing access. In the telecommunications context, this situation may be relevant for determining the access charges for the use of the incumbent's network for such services as mobile and many value-added services, the provision of which *might* be expected not to reduce demand for fixed-link voice telephony services significantly.

2.3.4 Discussion

This section has analyzed the case where the incumbent's retail tariff is fixed and where rivals must have a unit of access in order to provide a unit of their retail service. We showed that the optimal access charge in this situation is given by the sum of the cost of access and a "correction factor". From (7) this correction factor can be interpreted in two ways: it is the incumbent's lost profit in the retail sector caused by suppling access to its rivals as in (17), or it is the second-best output tax to correct for the fact that the incumbent's retail tariff does not reflect its costs as in (19). It is perhaps unfortunate that the former representation is more often used, as it gives entrants the excuse to complain that the ECPR acts to "maintain monopoly profits". (However, this complaint has much greater validity when the incumbent's retail price is not regulated—see section 2.6 below.) This view of the ECPR is extremely misleading. First, the analysis here is concerned with fixed retail tariffs, and if there is any "monopoly profit" present then this is the fault of regulation and not of the incumbent. Second, within the framework of this chapter the aim of the formula is *not* to recompense the incumbent for lost profit, but rather to give entrants the correct entry signals. For instance, if this opportunity cost element of the access charge were put into general public funds rather than paid to the incumbent, the efficiency aspects of the ECPR would be undiminished. (However, see footnote 27 above.)

Notice that an alternative way to implement the optimum would be to charge for access at the actual cost C_2 and at the same time to levy a second best tax on the *output* of rivals as proposed in section 2.1. However, in the present case where exactly one unit of access is needed to produce one unit of output, this output tax might just as well be levied on the input. When this fixed relationship between inputs and outputs ceases to hold, however, this convenient procedure cannot be applied. The next section discusses policy in this often more relevant situation.

2.4 Fixed Retail Prices and Bypass

Here we extend the previous analysis to allow for the possibility that entrants can substitute away from the network service offered by the incumbent, so that the demand for access (per unit of final output by the entrants) is a decreasing function of the access charge. We perform the analysis in two stages: first we examine the situation where the regulator has sufficient instruments to obtain the desired outcome; and secondly we discuss the choice of access charge given that this charge is the sole instrument of policy.

2.4.1 The Need to Price Access at Cost with Enough Instruments

In this section we assume that the regulator has enough instruments to implement the desirable outcome—see the discussion in section 2.2.3.[30] In particular, since the relationship between the entrant's inputs and outputs is not fixed, we suppose that the regulator can control both the price of access and the entrant's retail price. For instance, suppose that the regulator can levy an output tax on the entrants. (See the next section for an analysis of the case where only the former instrument is available.) We will concentrate on the unit demands framework as the competitive fringe model is so similar.

Unit demand model: Here we follow the bypass model outlined in section 2.2.1, assuming that the incumbent's retail price P is fixed by regulation. We wish to find a regulatory regime that ensures that the maximum value of welfare in (10) above is achieved. Specifically, suppose that E must pay the tax t per unit of its final output and the charge a per unit of M's network services. Following the earlier argument, given that E enters the market in some way, it will choose to use M's network if (11) holds. On the other hand, given that E enters in some way, (10) implies that welfare is higher when E uses M's network if

$$C_2 \leq [u - \hat{u}] + [\hat{C}_1 - c] \, .$$

Therefore, *given* that entry occurs, private and social incentives for using M's network are brought into line by choosing $a = C_2$. Making the network access charge equal to the cost of providing access gives the entrant the correct "make-or-buy" incentives for its network provision.[31]

[30] See also section 8 of Laffont and Tirole (1994) for related analysis. Section 3.2.4 of Laffont and Tirole (2000) discusses the benefits of levying output charges on entrants and notes that their use would imply that cost-based access charges are optimal. They regard these kinds of output charges as being "politically unlikely", however. They go on to suggest that these taxes could be repackaged as a tax on the whole industry to make them seem less discriminatory. This suggestion is illustrated in Table 3 in the current chapter.

[31] Several writers loyal to the ECPR approach have suggested that the ECPR is necessary for productive efficiency—see Baumol, Ordover, and Willig (1997) for instance. When bypass is possible, however, it is usually necessary to price access at cost, rather than at the ECPR level, to ensure productive efficiency at the network level.

Turning to the choice for t, following the analysis in section 2.1.1 the ideal output tax is given by $t = P - C_1$ per unit as in (3) above. With these choices for a and t we see that E's profits per unit with its three options for entry are:

$$E\text{'s profit } = \begin{cases} [\hat{u} - U] + [C_1 - \hat{C}_1] & \text{with stand-alone entry} \\ [u - U] + [C_1 - c - C_2] & \text{with entry via } M\text{'s network} \\ 0 & \text{with no entry .} \end{cases}$$

Just as was the case with M's profits in (12), comparing these profits with (10) we see that E's incentives are now precisely in line with welfare: the entrant will enter when it is optimal to do so, and will choose to use M's network when that is the more efficient mode of entry. Pricing access at cost means that the entrant has the correct make-or-buy incentives for network construction conditional upon entry, and the output tax (3) means that they have the correct incentives to enter (in any way) given that M's retail tariff is distorted. Other policies will cause various kinds of inefficiencies. For instance, if the entrant is permitted to use the incumbent's network at cost C_2 then it will face the correct make-or-buy incentives conditional on entry, but not the correct incentives to enter. Alternatively, if the ECPR charge (21) were imposed then the rival might build its own infrastructure even if it were more efficient for it to use the incumbent's.

As in section 2.1.1, the output tax element of this optimal policy can often be implemented by means of a suitably designed universal service fund, as described in Table 3.

	URBAN	**RURAL**
number of subscribers	20m	10m
M's cost per subscriber, of which	\$50	\$200
retail cost is	\$20	\$20
network cost is	\$30	\$180
M's retail price for service	\$100	\$100
M's profit for each type	\$1bn profit	\$1bn loss
Any firm's contribution to fund	\$50	–\$100
M's network access charge	\$30	\$180

Table 3: Giving correct entry and make-or-buy incentives

In this example there is a universal service fund that operates just as in Table 1: any firm providing service to an urban subscriber must contribute \$50 to this fund, and any firm offering service to a rural subscriber can receive \$100 from the fund. In addition to these contributions, the entrant can gain access to the incumbent's network at actual cost (not averaged costs as in Table 2). Notice that if the entrant chooses to enter via the incumbent's network its total payment is \$80 per subscriber, just as in Table 2. However, the advantage of splitting the "ECPR" charge into two parts—a cost-based access charge together with an

output tax—is that when network bypass is a possibility it is undesirable to make network access charges deviate from the incumbent's network costs.

Competitive fringe model: We discuss this briefly as it so similar to the unit demands case. Working with the model presented in section 2.2.2 above, suppose that the regulator fixed the access charge at a and the per-unit output tax on the fringe at t. Then, similarly to expression (13) above, total welfare in this case is

$$W = \underbrace{V(P, t+\psi(a))}_{\text{consumer surplus}} + \underbrace{(P - C_1)X(P, t+\psi(a))}_{M\text{'s profits from retail}}$$
$$+ \underbrace{(a - C_2)\psi'(a)x(P, t+\psi(a))}_{M\text{'s profits from access}} + \underbrace{tx(P, t+\psi(a))}_{\text{tax revenue from output tax}} \tag{24}$$

or writing $p = t + \psi(a)$ this simplifies to

$$W = V(P, p) + (P - C_1)X(P, p) + \{p - \psi(a) + (a - C_2)\psi'(a)\}\, x(P, p)\ . \tag{25}$$

Although this differs from the unregulated case in (14) by the addition of the consumer surplus term V, this extra factor does not affect the choice of a, which again is chosen to maximize the same term $\{\cdot\}$ and which again leads to marginal cost pricing of access: $a = C_2$. Finally, maximizing (25) with respect to $p = t + \psi(a)$ yields the formula (5) for t. Therefore, we see again that, when the regulator can use both these instruments, access should be priced at cost, and the entrants' output tax should be the "second-best output tax".

Thus we have obtained one of our main points in the chapter: just as in the unregulated case discussed in section 2.2 above, provided there are enough policy instruments available to pursue all the objectives, there is no need to sacrifice productive efficiency even when the incumbent's retail is not cost-reflective. Retail instruments—perhaps in the form of a carefully-designed universal service fund—should be used to combat retail-level distortions such as mandated tariffs that involve cross-subsidies. Wholesale instruments should then be used to combat potential productive inefficiencies—in this case the productive inefficiency caused by pricing access other than at cost.[32]

[32] This policy suggestion is somewhat related to the "M-ECPR" proposal as outlined in chapter 9 in Sidak and Spulber (1997a). Those authors suggest that the entrant should be charged an amount up to its *own* cost of providing network services for the use of the incumbent's network, and a "competitively neutral end-user charge" should be imposed to prevent cream-skimming entry. (See also Doane, Sibley, and Williams (1999) for further analysis.) One advantage, however, of basing access charges on the *incumbent's* cost is that it decentralizes the decision about the desirability of entry to the (perhaps better informed) entrant, and knowledge of the entrant's technology is not required.

2.4.2 Access Charges as the Sole Instrument: the ECPR Revisited

Although one of the main aims of this chapter is to argue that regulators should use output taxes for entrants—perhaps in the guise of a universal service fund—to correct for distortions in the incumbent's retail tariff, and use cost-based access charges to give the correct make-or-buy investment decisions, the former instrument is still only rarely used. Therefore, in this section we consider policy when the access charge is the sole instrument available to the regulator.[33] For simplicity, we discuss this issue in the context of the competitive fringe model of section 2.2.2 above.

In this case we impose $t = 0$ in (24), which yields welfare with the access charge a as

$$W = \underbrace{V(P, \psi(a))}_{\text{consumer surplus}} + \underbrace{(P - C_1)X(P, \psi(a))}_{M\text{'s profits from retail}} + \underbrace{(a - C_2)\psi'(a)x(P, \psi(a))}_{M\text{'s profits from access}} . \tag{26}$$

Maximizing this with respect to a gives

$$a = C_2 + \sigma(P - C_1) \tag{27}$$

where

$$\sigma = \frac{X_p \psi'(a)}{-z_a} \tag{28}$$

and $z(P, a) \equiv \psi'(a)x(P, \psi(a))$ is the fringe demand for access. This formula is again an instance of the ECPR formula (17), suitably interpreted. The first term of the right-hand side in (27) is the direct cost of providing access. The second term is the lost profit to the incumbent in the retail sector caused by providing the marginal unit of access to fringe. This lost profit is itself the product of two terms: M's marginal profit $(P - C_1)$ per unit of final product sales, and the *displacement ratio*, $\sigma = -X_p\psi'(a)/z_a > 0$. The parameter σ gives the reduction in demand for the incumbent's retail service caused by providing the fringe with the marginal unit of access (for a fixed retail price P).[34] Therefore, the second term in (27) indeed gives the loss in the incumbent's retail profit caused by supplying the marginal unit of access to the fringe.

This formula reduces to no-bypass rule (23) when $z \equiv x$, for in that case $\sigma = \sigma_d$. More generally, it is useful to decompose the displacement ratio σ into two terms as

$$\sigma = \sigma_d/\sigma_s ,$$

where $\sigma_d = -\frac{X_p}{x_p}$ as in (6) above, and

$$\sigma_s = \frac{z_a}{x_p\psi'(a)} = \frac{\psi''(a)x + (\psi'(a))^2 x_p}{x_p\psi'(a)} = \psi'(a) + \frac{-\psi''(a)\psi(a)}{\psi'(a)}\frac{1}{\eta_E} , \tag{29}$$

[33] This section is based on section III of Armstrong, Doyle, and Vickers (1996).

[34] For one more unit of access to be demanded by the fringe, the access charge has to fall by $1/z_a$, and this causes the demand for the incumbent's retail service to fall by $X_p\psi'/z_a$.

where $\eta_E = -px_p/x > 0$ is the own-price demand elasticity for fringe output. The term σ_d represents the effect of demand-side substitution possibilities, whereas σ_s represents the supply-side substitution possibilities. When there are no supply-side substitution possibilities, so that $\psi(a) = a + c$, then $\sigma_s \equiv 1$ and the displacement ratio is simply $\sigma = \sigma_d$. On the other hand, if the fringe output and the incumbent's output are perfect substitutes then $\sigma_d = 1$ and the displacement ratio is just $\sigma = 1/\sigma_s$.

Expression (29) shows that we may decompose σ_s itself into two terms: the first term $\psi'(a)$ captures the effect of changing a on the demand for M's access service caused by the change in fringe *output* (keeping the input mix constant), while the second term captures the effect of changing a on the demand for M's access service caused by changing the *input mix* (keeping fringe output constant). The term $\psi'(a)$ gives how many units of M's access service is needed for each unit of fringe output. The ability to substitute away from the incumbent's access service is captured by $-\psi''\psi/\psi'\eta$ in the above. Since this term is necessarily positive we immediately obtain the basic insight that the ability to substitute away from M's access service causes the displacement ratio to be reduced compared to the no-bypass benchmark (i.e. when $\psi'' \equiv 0$).

Notice that if rival services do substitute for the incumbent's retail service ($\sigma_d \neq 0$) then the ECPR rule in (27) implies that the access charge is not equal to the cost of access, which in turn implies that there is productive inefficiency whenever there is some scope for substitution ($\psi''(a) \neq 0$). The reason for this is that the access charge here is forced to perform two functions, and the regulator must compromise between productive and allocative efficiency. This problem is therefore analogous to that in the unregulated case covered in section 2.2.2.

2.5 Ramsey Pricing

Although the previous sections discussed the important problem of how to use access charges to maximize welfare for a *given* pattern of retail prices imposed on the incumbent, this approach leaves unexamined how these retail prices are chosen in the first place. Therefore, in this section we discuss the problem of optimally choosing the incumbent's retail and access prices simultaneously: the "Ramsey pricing" approach. For simplicity we discuss this problem only in the context of the competitive fringe model. As usual, the form of the solution will depend on whether or not bypass is an option, and, if it is, on the range of policy instruments available to the regulator.

2.5.1 No Bypass

Here we extend the model in section 2.3.3.[35] The problem is to maximize total welfare in

[35] This section is based on Laffont and Tirole (1994) and Armstrong, Doyle, and Vickers (1996). See also section 3.2 of Laffont and Tirole (2000).

(22) subject to the incumbent firm not running at a loss, i.e. that the variable profits—the final two terms in (22)—cover the fixed costs of the firm. Letting $\lambda \geq 0$ be the Lagrange multiplier associated with this constraint, we see that the retail price P and the access charge a are jointly chosen to maximize the following modification of (22):

$$W = V(P, c+a) + (1+\lambda)\left\{(a - C_2)x(P, c+a) + (P - C_1)X(P, c+a)\right\} \ .$$

Thus, there is now a greater weight placed on the incumbent's profit compared to the usual case, in order to reflect the need for charges to cover fixed costs. Writing $\theta = \lambda/(1+\lambda) \geq 0$, the respective first-order conditions for P and a are

$$P = C_1 + \frac{x_P}{-X_P}(a - C_2) + \frac{\theta P}{\eta_M} \tag{30}$$

where $\eta_M = -\frac{PX_P}{X} > 0$ is M's own price demand elasticity, and

$$a = \underbrace{C_2 + \sigma_d(P - C_1)}_{\text{ECPR access charge}} + \underbrace{\frac{\theta p}{\eta_E}}_{\text{Ramsey markup}} \tag{31}$$

where η_E is E's own price demand elasticity and σ_d is as in (6). Since the first two terms on the right-hand side of (31) replicate the corresponding ECPR formula in (23), this formula states that the optimal access charge is the ECPR level, which applies if P were exogenously fixed, plus a Ramsey markup. This Ramsey markup reflects the benefits—in terms of a reduction in P—caused by increasing the revenue generated by selling access to the fringe. In particular, the Ramsey access charge is *above* the ECPR recommendation (which applies taking as given the retail price P). The reason for this is that a higher a raises more revenue that can be used partly to cover the fixed costs, and this allows P to be lowered (which is good for welfare).[36]

It is useful to compare these expressions with the alternative expressions (13) and (18) in Laffont and Tirole (1994), which state (using this chapter's notation) that

$$P = C_1 + \frac{\theta P}{\hat{\eta}_M} \ , \ a = C_2 + \frac{\theta p}{\hat{\eta}_E}$$

where $\hat{\eta}_M$ and $\hat{\eta}_E$ are the *superelasticities* of the respective products.[37] Although these two pairs of equations look very different, it is possible to show that the expressions (30)

[36] Section 8 of Laffont and Tirole (1996) and section 4.7 of Laffont and Tirole (2000) make the important policy point that Ramsey prices can be implemented by means of a 'global price cap', where both the access and retail services of the incumbent are controlled by means of a suitably designed average price cap. As well as treating wholesale and retail services more symmetrically, they argue that this regulatory mechanism gives the incumbent fewer incentives to exclude rivals by non-price means (compared to, say, cost-based access charges).

[37] The reason that the two pairs of first-order conditions look different is that this chapter's are obtained by maximizing over prices, whereas Laffont and Tirole maximize over quantities.

and (31) are indeed equivalent to those of Laffont and Tirole. In particular, Laffont and Tirole's expression for the optimal access charge which expresses a as a markup over the cost of providing access involving the superelasticity, may be re-expressed as we have done in (31), where a is expressed as a markup over the ECPR level involving just the *normal* elasticity—see section 3.2.2 of Laffont and Tirole (2000) for a similar analysis.

An important issue is the relationship between the Ramsey approach to access pricing and the ECPR approach. Obviously, this question hinges on what we mean by the term "ECPR"—see section 2.3.1 for various interpretations of this rule. Using our preferred interpretation, which is (23), we see that Ramsey pricing never leads to an ECPR access charge. If, however, one takes the margin rule (21) as the appropriate benchmark then it is indeed possible for the Ramsey access charge to happen to coincide with this margin rule. This issue is discussed in section 3.2.5 of Laffont and Tirole (2000), who show that with enough symmetry between the incumbent and the fringe the Ramsey access charge does satisfy (21), where P is endogenously determined by the Ramsey problem. However, it is not clear why the margin rule should be a relevant benchmark in this context: with product differentiation the correct opportunity cost incurred by the incumbent in providing a unit of access service is given by (23) and not by (21).

2.5.2 Bypass

Here we extend the Ramsey analysis to allow for bypass by entrants, as in section 2.2.2. Suppose first that the fringe pays a per-unit output tax equal to t (as well as the per-unit charge a for access input). As usual, the price of the fringe product is equal to the perceived marginal cost, so that $p = t + \psi(a)$, and fringe profits are zero.

As in 2.2.2 and 2.4.1 above, the regulator can be considered to choose p directly rather than t, in which case the incumbent's profit Π is (14). Let $\lambda \geq 0$ be the shadow price on the profit constraint $\Pi \geq F$, where F is the fixed cost that needs to be covered by profits. Then the problem is to choose P, p and a to maximize $W = V(P, p) + (1 + \lambda)\Pi$. However, since for given retail prices P and p the access charge a does not affect consumer surplus, it is clear that a must be chosen to maximize Π for given retail prices, so that a again maximizes the term $\{\cdot\}$ in (14). As before, provided there are some possibilities for substituting away from the incumbent's access product, the first-order condition for this cost minimization is $a = C_2$ and so pricing access at cost is optimal.[38] This is just an instance of the deep result

[38]This paragraph provides one argument, to do with distortions at the input level, for the use of both retail and wholesale instruments for policy. Another rationale for this might be the following: Ramsey principles imply that different retail services that use the same access service should typically have different retail prices, depending on the demand elasticities. When there is no scope for bypass, in some cases one could implement this outcome by differential, use-dependent access charges. In others, though, this may not be possible, perhaps because the incumbent cannot accurately monitor the use to which its network is put. In such circumstances differential retail taxes could be used to implement the Ramsey solution, and non-discriminatory cost-based charges could then be used for network access.

that productive efficiency is desirable when there are enough tax instruments—see Diamond and Mirrlees (1971).[39]

Next, as in section 2.4.2, suppose the regulator has a more limited set of policy instruments, and that the output tax t is not available. In this case $p \equiv \psi(a)$ and the access charge must perform two functions: it must attempt to maintain productive efficiency (as before) but in addition it must influence the fringe retail price in a desirable way. The incumbent's profit is now as in (15). Writing $\theta = \lambda/(1+\lambda) \geq 0$, the respective first-order conditions for maximizing $V + (1+\lambda)\Pi$ for P and a are respectively

$$P = C_1 + (a - C_2)\frac{z_P}{-X_P} + \frac{\theta P}{\eta_I}$$

and

$$a = \underbrace{C_2 + \sigma(P - C_1)}_{\text{ECPR access charge}} + \underbrace{\frac{\theta a}{\eta_z}}_{\text{Ramsey markup}}, \tag{32}$$

where σ is as given in (28), and $\eta_z = -\frac{a z_a}{z}$ is the own price elasticity of the demand for access. In particular, the Ramsey access charge is again above the associated ECPR recommendation, which this time is given by (27). These first-order conditions imply that $P > C_1$ and $a > C_2$, and so access is priced above marginal cost. This in turn leads to a degree of productive inefficiency. Just as in sections 2.2.3 and 2.4.2 the access charge is called upon to perform too many tasks, and a compromise must be made. In the next section the access charge is forced to perform yet another task, which is to try to control the *incumbent's* unregulated retail price.

2.6 Unregulated Retail Prices

In this section we discuss how best to price access when this is the only instrument for regulatory control of the incumbent.[40] In particular, the incumbent is now assumed to be free to set its retail price. For simplicity we suppose that there is no output tax on the fringe. (It would seem strange to consider a situation where the entrants were regulated in some sense while the incumbent was not.)

[39] On pages 370–71 of Sidak and Spulber (1997a) those authors try to argue that pricing access differently from cost—as with ECPR-style access charges—does not violate the Diamond-Mirrlees result.

[40] This is adapted from section 7 of Laffont and Tirole (1994), section 5.2.2 of Armstrong, Cowan, and Vickers (1994) and Armstrong and Vickers (1998). For other analyses of access pricing with an unregulated downstream sector, see Lewis and Sappington (1999), Economides and White (1995), Lapuerta and Tye (1999) and Vickers (1995). In addition, sections 3 and 4 of this chapter consider other situations where firms are unregulated at the retail level.

2.6.1 Perfect Retail Competition

To discuss this topic in its simplest and starkest form, we first model downstream competition as being perfect. Specifically, there is a group of rivals which offers a service which is a perfect substitute for the incumbent's. Consumer demand for this homogeneous service is denoted $Q(P)$, where P is the price offered by firms in the market. For simplicity suppose the rivals cannot bypass the incumbent's access service, and it costs rivals c to convert a unit of access into a unit of retail product.[41] As usual, the incumbent has marginal cost C_1 for supplying its end-to-end retail service and cost C_2 for providing access to the fringe. Therefore, it is efficient for the fringe to supply the market whenever

$$c + C_2 \leq C_1 \ . \tag{33}$$

Control of the access charge: Suppose regulation fixes the access charge at a. Then there will be entry if M chooses its retail price above $a+c$. In this case M will get profit $a-C_2$ per unit from providing access to the rivals. On the other hand, if M wishes to stay in the retail market the maximum retail price it can charge is the limit price $P = a + c$, which generates profit $a + c - C_1$ per unit. Therefore, the incumbent will choose to allow entry if and only if this is efficient as in (33). (The level of the regulated access charge does not affect the relative profitability of the two strategies, although it will affect the absolute profitability of either.) We deduce that, when the incumbent is free to set its retail price, it will allow entry if and only if the entrant has lower costs, regardless of the level of the access charge. In particular, because of opportunity cost considerations, the fact that a differs from C_2 does not distort at all the incumbent's incentives to compete on a "level playing field" with its downstream rivals.

Since we have seen that productive efficiency is automatically ensured in this perfect competition setting, the access charge should be chosen to attain allocative efficiency. Notice that, regardless of whether entry is successful, the equilibrium retail price given a is $P = a+c$ (at least if this is below the unregulated monopoly retail price). Therefore, we wish to choose a so that this price equals (minimized) marginal cost, so that $a + c = \min\{C_1, C_2 + c\}$. In the case where entry is more efficient this implies that $a = C_2$ and access is priced at marginal cost. In the case where the incumbent should serve the market, we should ideally set $a = C_1 - c$.[42] However, when the market is fairly symmetric, in the sense that $C_1 \approx C_2 + c$, pricing access at marginal cost will be a good approximation to the ideal access pricing regime.[43]

[41] Little would change if we allowed for the possibility of bypass, since this would just strengthen the motive to choose an access charge close to cost.

[42] By construction, this access charge is below the cost of providing access, C_2, in order to eliminate the mark-up that the more efficient incumbent would otherwise be able to charge in this market.

[43] This is essentially a formalization of the argument in section 7 of Lapuerta and Tye (1999), who argue that access should be priced at cost because competition in the retail sector can be relied upon to eliminate all distortions in that segment.

Control of the margin: Next, consider the effects of controlling the margin $P - a$ rather than the access charge a, so that M can choose any pair of prices P and a that satisfy $P - a = m$, say. Clearly, entry takes place if and only if this regulated margin covers the fringe's retail cost, i.e. if $m \geq c$. In particular, with this regime M has no discretion about whether or not entry takes place. If entry does take place then, given the access charge a, the fringe retail price will be $p = a + c$, and M's profit will be $(a - C_2)Q(a + c)$. If we think of M as choosing the fringe retail price $p = a + c$ rather than a, then p will be chosen to maximize $(p - c - C_2)Q(p)$. On the other hand, if $m < c$ then M will choose its retail price P to maximize its profits $(P - C_1)Q(P)$. In either case we have the outcome corresponding to unregulated monopoly—with entry the marginal cost is $c + C_2$ and with the incumbent serving the retail market the marginal cost is C_1—and this kind of regulation exerts no downward pressure on retail prices whatsoever.[44] However, since welfare with unregulated monopoly increases when the marginal cost is reduced, given that margin regulation is being used, it is better to have entry if and only if it is efficient. Therefore, we want m to be chosen so that $c \leq m$ if and only if $c + C_2 \leq C_1$. In other words, optimal margin regulation entails $m = C_1 - C_2$, i.e. that the margin rule (21) should be applied.

However, it is simple to verify that this is exactly the outcome when M is totally unregulated. (In section 2.2.1 we saw in a similar model that the unregulated incumbent will always allow entry when this is efficient, and will then maximize profits.) Thus we see that the *best* form of margin regulation—which is given by the margin rule version of the ECPR—simply replicates the totally unregulated outcome. This provides a compelling argument against the use of the ECPR in deregulated markets. Moreover, this argument gives validity to the common complaint that the ECPR acts to "maintain monopoly profits".[45] (But see the discussion in section 2.3.4 for why this complaint is not valid when the incumbent's retail tariff is controlled by regulation.) This policy contrasts with the preceding case where the access charge is the instrument of policy, and where a low access charge translates directly into low retail prices. Therefore, we can deduce that direct control of the access charge is superior in terms of welfare compared to margin regulation. Indeed, in this perfect competition framework, pricing access at cost—or a little lower in some cases, if feasible—is then the optimal policy.

2.6.2 Competitive Fringe Model

To investigate this topic in more detail, we return to the competitive fringe model with bypass as in section 2.2.2. Therefore, the fringe equilibrium retail price is $p = \psi(a)$, and

[44] This has a similar flavor to the duopoly analysis in pages 26–27 of Laffont, Rey, and Tirole (1998a). Those authors find that a form of margin regulation—where networks choose their access charge and then the allowed retail prices must generate a specified margin—facilitates collusion, i.e. that there is again no downward pressure on retail prices.

[45] This point is forcefully made in section II(A) of Tye and Lapuerta (1996).

M's profit as a function of P and the access charge a is as given in (15) above. For a given access charge a, the incumbent chooses P to maximize its profits, the solution to which has the first-order condition (16). Note that this may be rearranged to give

$$a = C_2 + \underbrace{\frac{-\Pi_P^R}{z_P}}_{M\text{'s lost retail profit}} . \tag{34}$$

This can again can be interpreted as an instance of the ECPR rule (17) above.[46] The first term on the right-hand side of the above is the cost of providing access, while the second is M's loss in profits in the retail sector caused by supplying the marginal unit of access to the fringe.[47] In addition, since $z_P > 0$ (34) implies that Π_P^R has the opposite sign to $(a - C_2)$. This has a natural intuition: if access is priced above cost, the incumbent has an incentive to push its retail price *above* the profit-maximizing level for the retail sector viewed in isolation since it also wishes to stimulate demand for its profitable access service. (The reverse argument holds if $a < C_2$.)

The profit-maximizing choice of P given a, denoted $\bar{P}(a)$, maximizes Π in (15). In most reasonable case it makes sense to assume that $\bar{P}'(a) > 0$, so that a higher regulated access charge leads to a higher equilibrium retail price for the incumbent: the more profitable selling access to its rivals is, the less aggressively the incumbent will compete with rivals. (The following analysis can easily be adapted if this assumption is invalid.) The welfare-maximizing choice for a is derived as follows. Let $\bar{\Pi}(a)$ be the incumbent's maximum profit in (15) given a. Choosing a to maximize welfare $V(\bar{P}(a), \psi(a)) + \bar{\Pi}(a)$ implies that

$$-X\bar{P}' - x\psi' + \bar{\Pi}' = 0 ,$$

which, by using the envelope theorem for $\bar{\Pi}'$ and the fact that $z = x\psi'$, implies that

$$a = \underbrace{C_2 + \sigma(\bar{P}(a) - C_1)}_{\text{ECPR access charge}} - \underbrace{\frac{X\bar{P}'}{-z_a}}_{\text{mark-down to control } P} \tag{35}$$

where σ is given in (28). Therefore the access charge should be *below* the ECPR level in (27) given by the situation where M's retail price was *fixed* at $\bar{P}(a)$: the fact that M's retail price is unregulated implies that the access charge should be set below the ECPR level that should apply if its retail price were fixed. The intuition for this is clear. If controlling M's

[46] One would not want to push this interpretation too far. In sections 2.3 and 2.4 we made the normative point that the optimal access charge given a specific and fixed retail price was the ECPR. In this section we merely make the *positive* point that, for any given access charge, the unregulated incumbent's choice of retail price happens to have an ECPR flavor.

[47] To cause a further unit of access to be demanded requires, for a given a, that P rises by an amount $1/z_P$, which in turn causes profits in the retail sector to fall by the amount given in the formula.

retail price were not an issue, section 2.4.2 argued that the optimal access charge was given by the ECPR in (27). But now a reduction in a causes the retail price P to fall, which is good for welfare. In sum, because M's retail price is positively related to its access charge, there is a need to reduce the access charge to below the ECPR level.[48] (By contrast, in the Ramsey problem it was optimal to raise the access charge from the ECPR level—see expression (31) above—for the reason that an *increase* in the access charge there caused the retail price to fall, since the access service then financed more of the fixed costs of the firm.)

Another natural comparison is between a and the cost of access C_2. (The fact that a is below the ECPR level says little about the comparison of a with C_2.) In the previous section with perfect competition, we saw that pricing access at cost—or sometimes a little below cost if this were feasible—was the optimal policy. However, in this more general framework it seems hard to obtain clear-cut results about whether a should be above or below cost, and in general either can be optimal. However, in a few special cases the optimal access charge should precisely equal cost:

- With no possibility for bypass and if the demand functions X and x are linear then the regulator should set $a = C_2$.[49]
- If there are no cross-market effects then the regulator should set $a = C_2$. (From (15) the profit-maximizing retail price $\bar{P}$ does not depend on a in this case, and also $\sigma = 0$, therefore (35) implies that marginal cost pricing is optimal.)

However, in general it will simply be by chance that the optimal access charge is equal to cost. As a result we expect that when bypass is a possibility the result will be productive inefficiency. The reason that it is hard to get clear-cut results in this framework is because the access charge here is called upon to perform *three* tasks: (i) it is used to control the market power of the incumbent (a lower value of a feeds through into a lower value for P); (ii) it is used to achieve allocative efficiency *given* P using the second best argument of section 2.3.3, and (iii) it is used to try to achieve productive efficiency (which requires $a = C_2$) whenever there is a possibility for bypass. In general, motives (i) and (iii) argue for an access charge no higher than cost. (When $a = C_2$ the incumbent will choose $P > C_1$, and so choosing $a < C_2$ will bring M's retail price down towards cost. Motive (iii) will merely temper but not overturn this incentive.) However, unless a is chosen to be so low that $P < C_1$, motive (ii) will give the regulator an incentive to raise a above cost—see the ECPR expression (27). Therefore, because of these forces pulling in different directions, it is not possible to give clear guidance about the relationship between the access charge and the cost of providing access in unregulated retail markets.

[48] A similar point is made in section III of Economides and White (1995). They show that when the downstream market is unregulated it can be desirable to allow entry by an inefficient firm—something that is achieved by choosing an access charge below the ECPR level—if this causes retail prices to fall, i.e. it can be a good thing to sacrifice a little productive efficiency to reduce allocative inefficiency.

[49] See section 7 of Laffont and Tirole (1994) and Armstrong and Vickers (1998).

2.6.3 Partial Deregulation

In practice the incumbent firm often operates in retail markets that are partially regulated.[50] Thus the firm may have discretion choosing its retail price in a given market, but only subject to certain constraints on the overall pattern of its retail prices. There are two main kinds of constraint that are commonly imposed. First, the incumbent's retail tariff could be controlled by some kind of average price cap, so that it has freedom to vary relative retail prices subject to an overall cap. In this case a decrease in one retail price—say, in response to entry—then allows the firm to increase other retail prices. Second, the firm may face constraints on "price discrimination", so that it is free to determine the *level* of its retail prices but faces constraints on the structure of relative prices (such as a geographically uniform tariff restriction). Here, if the firm decreases one price in response to entry then it is forced to *decrease* other prices in related markets.

How is our analysis modified by these cross-market interactions? Suppose for simplicity there are two retail markets, in one of which the incumbent faces entry and in the other no threat from entry. Suppose also that consumer demand in the two markets is independent. Write P_1 to be M's price in the potentially more competitive market, and let P_2 be the price in its captive market. As usual, let a be the network access charge. In this case the firm's profit in (15) is modified to be

$$\Pi = \underbrace{\Pi_2^R(P_2)}_{\text{captive market}} + \underbrace{\Pi_1^R(P_1,a)}_{\text{competitive market}} + \underbrace{(a-C_2)z(P_1,a)}_{\text{access market}} \tag{36}$$

where $\Pi_1^R(P_1,a) \equiv (P_1-C_1)X_1(P_1,\psi(a))$ is M's profit in the more competitive retail market, $\Pi_2^R(P_2)$ is its profit function in the captive market, and $z(P_1,a) \equiv \psi'(a)x(P_1,\psi(a))$ is the fringe demand for the access needed for its activity in the competitive market.

Suppose that M is constrained by policy to choose $P_2 = \zeta(P_1)$. When the firm faces an average price cap, the relationship ζ is decreasing: a rise in P_1 requires a fall in P_2. By contrast, a ban on price discrimination across the two markets could be represented by the constraint $P_2 = P_1$, so that ζ is increasing. If we impose the constraint $P_2 = \zeta(P_1)$ in (36), write $\bar{P}_1(a)$ to be the resulting profit-maximizing price P_1 when the access charge is a. In most reasonable cases we continue to have $\bar{P}_1(a)$ being an increasing function of the access charge: the more profitable selling access to its rivals is, the less aggressively the incumbent will compete when faced with entry. Then (35) is modified to be

$$a = \underbrace{C_2 + \sigma(\bar{P}_1(a) - C_1)}_{\text{ECPR access charge}} - \underbrace{\frac{(X_1 + X_2\zeta')\bar{P}_1'}{-z_a}}_{\text{correction to control } P_1,P_2}. \tag{37}$$

[50] See section 5 in Vickers (1997) and section 4.7 in Laffont and Tirole (2000) for some discussion of this topic. This topic is also closely related to the analysis of Armstrong and Vickers (1993), although there are no vertical issues in that paper.

(Here $\zeta' = \zeta'(\bar{P}_1(a))$ and X_1 and X_2 are the incumbent's equilibrium demands in its two retail markets.)

The effect of the cross-market price constraints are then as follows. First there is an effect on the firm's incentive to compete, as captured by $\bar{P}_1$ and $\bar{P}_1'$. In the case of a ban on price discrimination we expect that, because the firm is reluctant to lose profits in its captive market, it is less responsive to changes in the access charge. Indeed, in the limit as profits in the captive market become the incumbent's sole objective then $\bar{P}_1' = 0$ and (37) implies that the usual ECPR formula is then optimal. In effect, the retail price in the competitive market is now fixed exogenously, and the earlier analysis of access pricing with fixed retail prices in section 2.4.2 goes through as before. In less extreme cases the presence of the captive market will just temper the incumbent's incentive to vary P_1 in response to policy towards the access charge, and the correction factor on the right-hand side of (37) will correspondingly be reduced.

Secondly, when the firm operates under an average price cap, then typically this correction factor will again be reduced, and the ECPR formula will again be closer to being optimal than indicated in section 2.4.2. For instance, suppose that the price cap takes the "fixed weights" form: the incumbent must choose retail prices satisfying $w_1 P_1 + w_2 P_2 \leq P^*$ for some positive constants w_1 and w_2. In this case $\zeta' \equiv -w_1/w_2$. Suppose further that these fixed weights are chosen to be proportional to the equilibrium demands in the two markets, so that $w_1/w_2 = X_1/X_2$.[51] In this case the correction factor in (37) vanishes entirely, and the ECPR is again valid. In this special case the captive market provides exactly the appropriate correction factor to control the incumbent's retail price in the competitive market, and the access charge can be chosen as if the retail price in the more competitive market were fixed.

2.7 Introducing Dynamic Issues

Until this point (as indeed it will be for the remainder of the chapter) the analysis has been purely static. To see how this static analysis generalizes easily to encompass some dynamic aspects, consider this simple extension to the no bypass model of section 2.3.3.[52] Investment, production and consumption take place at discrete points in time, $t = 0, 1, ...$, and suppose that the prices for M's product and E's product in period t are, respectively, P_t and p_t. Suppose that consumer surplus, and demand functions for the incumbent's and fringe's products in period t are, respectively, $V_t(P_t, p_t)$, $X_t(P_t, p_t)$ and $x_t(P_t, p_t)$. (We assume there

[51] Section 2.2.2 of Laffont and Tirole (2000) discusses this form of average price regulation in more detail. In particular, they show that, in the absence of competition, this form of price regulation leads to Ramsey-like prices.

[52] The following extends the analysis in section 4.4.1.3 of Laffont and Tirole (2000). See Hausman and Sidak (1999) and Jorde, Sidak, and Teece (2000) for analyses of the important issue of how dynamic access price regulation affects the incumbent's incentive to innovate and upgrade its network. For a full treatment of dynamic issues in access pricing which emphasizes the problem of commitment and expropriation in regulatory policy, see Sidak and Spulber (1996), Sidak and Spulber (1997a) and Sidak and Spulber (1997b).

are no intertemporal linkages in demand.) Over time the incumbent invests in a network which supplies access, and suppose that one unit of access is needed to provide one unit of retail service, either by M or by E. Suppose that the capacity of the network at time t, which determines the total number of units of retail service that can be generated by the network, is K_t. Capacity depreciates at the proportional rate δ. There are constant returns to scale in installing capacity, and suppose that installing a unit of capacity in period t costs β_t. Let I_t be the amount of investment (in money terms) in period t, so that the amount of new capacity installed in period t is I_t/β_t. Then capacity evolves according to the dynamic relation

$$K_{t+1} = (1-\delta)K_t + \frac{I_{t+1}}{\beta_{t+1}} \ . \tag{38}$$

(We suppose that the current period's investment can be used with immediate effect.)

What is the marginal cost of providing an extra unit of access in period t? Suppose the investment plan is $K_t, K_{t+1}, ..., I_t, I_{t+1}, ...$ satisfying (38). If K_t is increased by 1, then all subsequent values for K and I are unchanged provided that next period's investment I_{t+1} is reduced so as to the keep the right-hand side of (38) constant, i.e. if I_{t+1} is reduced by $(1-\delta)\beta_{t+1}$.[53] If the interest rate is r, so that \$1 next period is worth $\$\frac{1}{1+r}$ now, then the net cost of this modification to the investment plan is

$$LRIC_t = \beta_t - \frac{1-\delta}{1+r}\beta_{t+1} \ .$$

(The notation $LRIC$ is used to stand for 'long-run incremental cost', the term often used in policy debates.) If technical progress causes the unit cost of new capacity to fall at the exogenous rate γ every period, then $\beta_{t+1} = (1-\gamma)\beta_t$. With technical progress γ, the above formula reduces to

$$LRIC_t = \beta_t \left(1 - \frac{(1-\delta)(1-\gamma)}{1+r}\right) \ . \tag{39}$$

Notice that if the parameters δ, r and γ are each reasonably small, this formula is approximated by $LRIC_t \approx \beta_t(r+\gamma+\delta)$.[54]

Suppose that it costs M and E an amount C_t and c_t respectively to convert a unit of access into their final retail services. Let a_t be the access charge paid by E in period t. Then competition implies that the fringe's price in period t is $p_t = a_t + c_t$, and total discounted welfare is

$$W = \sum_t \frac{1}{(1+r)^t} \left\{V_t + X_t(P_t - C_t) + x_t a_t - I_t\right\} \ ,$$

[53] I assume that demand conditions are such that investment in each period is strictly positive, which ensures that this modification is feasible.

[54] This is a familar equation in continuous time investment models—see for instance expression (7) in Biglaiser and Riordan (2000).

where $K_t \equiv X_t + x_t$ and this capital stock evolves according to (38). (All functions in the above are evaluated at the prices P_t and $p_t = a_t + c_t$.) Since (38) implies that

$$I_t = \beta_t \left[X_t + x_t - (1-\delta)(X_{t-1} + x_{t-1}) \right] \ ,$$

substituting this value for investment into the expression for welfare W yields

$$W = \sum_t \frac{1}{(1+r)^t} \left\{ V_t + X_t(P_t - C_t) + x_t a_t - \beta_t \left[X_t + x_t - (1-\delta)(X_{t-1} + x_{t-1}) \right] \right\} \ .$$

Maximizing this with respect to P_t and a_t yields

$$\left\{ \frac{\partial X_t}{\partial P_t}(P_t - C_t) + \frac{\partial x_t}{\partial P_t} a_t - \beta_t \frac{\partial [X_t + x_t]}{\partial P_t} \right\} + \frac{1}{1+r} \left\{ (1-\delta)\beta_{t+1} \frac{\partial [X_t + x_t]}{\partial P_t} \right\} = 0$$

$$\left\{ \frac{\partial X_t}{\partial p_t}(P_t - C_t) + \frac{\partial x_t}{\partial p_t} a_t - \beta_t \frac{\partial [X_t + x_t]}{\partial p_t} \right\} + \frac{1}{1+r} \left\{ (1-\delta)\beta_{t+1} \frac{\partial [X_t + x_t]}{\partial p_t} \right\} = 0 \ . \qquad (40)$$

Together these imply that the access charge should be given by $a_t = LRIC_t$ in (39), and that $P_t = C_t + a_t$. Thus, as expected in this constant-returns-to-scale world with all prices chosen to maximize welfare, it is optimal to set all charges equal to the relevant marginal costs. In particularly, it is optimal at each point in time to set a_t equal to the correctly calculated marginal cost (which falls at the rate of technical progress γ each period).

In order to bring out the similarities with section 2.3, we next calculate optimal access charges $\{a_t\}$ for a *given* time-path for the incumbent's retail prices $\{P_t\}$. These are derived from the single expression (40) above, which can be simplified. As in (6), if we write

$$\sigma_t = \frac{\partial X_t / \partial p_t}{-\partial x_t / \partial p_t}$$

for the demand substitution parameter in period t, then (40) simplifies to

$$a_t = \underbrace{LRIC_t}_{\text{access cost in } t} + \underbrace{\sigma_t \left[P_t - C_t - LRIC_t \right]}_{\text{lost retail profit in } t}$$

where $LRIC_t$ is given in (39). This is just a dynamic version of the ECPR rule (23) derived above. Thus we see that the formulae generated in the simple static models extend in a straightforward way to simple dynamic models.[55]

[55] For further analysis of the dynamic access pricing problem, focussing on the issue of how to encourage multiple networks to provide infrastructure investments at the efficient level, see Gans and Williams (1999) and Gans (2001).

2.8 Controversies and Conclusions

It is hard to find a more controversial issue in industrial policy than that concerning the terms on which entrants can gain access to an incumbent firm's network. In this section I try to summarize the main insights generated by the analysis, and how these can shed light on the current policy debates.

Three broad kinds of access charging policy are:

1. Pricing access at cost,
2. Ramsey pricing, i.e. choosing the incumbent's retail prices and access charges simultaneously to maximize welfare, and
3. the ECPR, i.e. pricing access at the cost of access plus the incumbent's opportunity cost.

Economic opinion on the applicability of these policies often seems to be extremely polarized, and cost-based access charges and ECPR-based access charges both have their fervent supporters. As usual, though, the relative merits of these policies depends on the specifics of each case, and these are discussed in turn.

2.8.1 Cost-Based Access Charges

The chief benefits of cost-based access charges are twofold. First, they are relatively simple to implement. (Or at least as simple as estimating the incumbent's network costs, something which is needed for all reasonable access pricing policies.) In particular, to calculate these charges no information is needed about subscriber demand, nor about the characteristics of entrants (at least in the simple models presented above). Second, this is the only access pricing policy that gives the correct "make-or-buy" signals to entrants when bypass is a possibility. For instance, pricing access above cost could mean that an entrant would prefer to bypass the incumbent's network and construct its own network, even though it would be more efficient to use the incumbent's network. A third, and less clear-cut, benefit of such charges is that they are "fair and non-discriminatory", and do not depend upon the use which is made of the incumbent's network by rivals. Therefore, under this regime different entrants will not be offered different wholesale terms by the incumbent.

In simple terms, cost based access charges are appropriate when access charges do not need to perform the additional role of correcting for distortions in the incumbent's retail tariff. There are three main reasons why such a task might not be necessary:

1. First, if the incumbent's regulated retail tariff *does* reflect its underlying costs accurately then no "second-best" corrective measures are needed at all, and in such cases access charges should also reflect the relevant costs. In sum, a full and effective rebalancing of the incumbent's tariff greatly simplifies the regulatory task, and allows access charges to perform the focussed task of ensuring productive efficiency.

2. Second, if there are distortions present in the regulated tariff, but the second-best corrections are made via another regulatory instrument (such as an output tax levied on entrants), then access charges need not perform this additional task, and so again can safely reflect costs—see sections 2.4.1 and 2.5.2 above. Indeed, one of the main aims of this analysis has been to argue that cost-based access pricing *is* the best policy, provided that the incumbent's retail distortions are more directly tackled by, for instance, a well-designed universal service fund.

3. Third, when the incumbent operates in a vigorously competitive retail market and is free to set its own retail tariff, we argued that pricing access at cost was a good policy—see section 2.6.1 above. The fact that the downstream sector is competitive implies that the incumbent has no significant opportunity costs, and so once again access charges should reflect costs.

However, in other cases—i.e. when opportunity cost considerations apply and when access charges are required to correct for these—we have seen that pricing access at cost is sub-optimal.

2.8.2 Ramsey Pricing

Almost by definition, Ramsey pricing is the best way to set access (and other) prices. Once the regulator has chosen a measure of social welfare—which could include special care being taken over the welfare of certain subscriber groups—then the optimal policy is to choose access and retail charges to maximize this welfare function, subject to constraints on the profitability of the firm and/or the costs of public funds. This is clearly superior to a policy whereby retail prices are (somehow) chosen first, perhaps without regard for how access charges will subsequently be chosen, and access charges are then chosen taking this retail tariff as given. Nevertheless, it seems fair to say that in practice Ramsey pricing principles are not often heeded for regulated retail tariffs, and access charges are left to correct for the various resulting retail distortions. One possible reason for this is that retail tariffs are much more "visible" than access charges, and decision makers are more susceptible to public and political pressure when they choose these. (This is not to deny that firms in the industry are a powerful lobbying force when it comes to influencing the access charging regime.)

A common argument against the use of Ramsey prices is that they are very informationally demanding, and that compared to, say, cost-based prices, they require knowledge about demand elasticities and so on, which the regulator simply does not have. While it is broadly true that regulatory bodies do not in fact possess this kind of detailed market information, this is not to say that with further effort they could not obtain reasonable estimates of these data. Alternatively, ways could perhaps be found to delegate pricing decisions to the firm, which will most likely be much better informed than the regulator about the market in which it operates. The global price cap proposed in section 4.7 of Laffont and Tirole (2000) provides one framework in which to do this.

Another, more dubious, argument is that Ramsey-style access charges are "discriminatory", in the sense that the charge a rival must pay for a given access service will depend on the use to which this service is put. Our formula (32), for instance, shows that, as well as the cost of providing access, the access charge should depend on (i) the incumbent's (endogenous) price-cost margin in the relevant retail market, (ii) the displacement ratio σ, which takes account of demand-side substitution possibilities as well as supply-side bypass possibilities, and (iii) the elasticity of demand for access. Each of these factors will depend on the particular use made of the access service. But the point is that this is *desirable*: Ramsey charges are the "least-bad" departures from cost-based charges, and access charges should be higher when used for those services that do not generate large welfare losses when significant price-cost markups are imposed.

It is fair to say the Ramsey approach does not enjoy the same passionate level of support or disparagement from economists as do the other two policies. One might speculate that this is because the policy is not strongly supported by either incumbents or entrants in the industry. (Very roughly, cost-based access pricing is supported by entrants, whereas the ECPR, in its simple forms at least, is supported by incumbents.) In regulatory hearings around the world, these firms find economists to support their respective cases, but there is no well-funded entity that argues strongly for Ramsey-style pricing of incumbent services.

2.8.3 The ECPR

The ECPR must be one of the more misunderstood formulas in industrial economics—see section 2.3.1 for a variety of interpretations. The analysis in this section has argued, broadly speaking, that with our preferred version of the rule as given by (17) the rule is valid when (a) the incumbent's retail tariff is fixed in advance and is not affected by any actions of the rivals, and (b) when other instruments, such as an output tax levied on rivals, are not available to correct for the incumbent's retail market distortions. In particular, the rule has little relevance when the incumbent is free to choose its retail tariff, and in some cases it does nothing to constrain incumbent monopoly power—see section 2.6. (However, with *partial* deregulation the rule tends to perform better.)

In the guise of the margin rule (20) it has the virtue of simplicity and being informationally undemanding (at least when compared to the Ramsey approach). All that needs to be known is the incumbent's retail price and its avoidable costs in the retail segment. However, this margin rule is simply not appropriate except in a few special cases, such as the one discussed in section 2.3.2 for instance. If the margin rule is used as a basis for policy then inefficiencies will result in situations where (i) the rival's product does not substitute one-for-one for the incumbent's, or (ii) where rivals have some ability to bypass the incumbent's access service.

On the other hand, our preferred version of the ECPR, as represented by (17), is, outside these same special cases, informationally demanding, and various awkward elasticities appear in our expressions for opportunity cost. This suggests that the apparent contrast between the "simple" ECPR approach and the "complex" and informationally demanding Ramsey

approach may just be a misleading artefact of simple examples with extreme elasticities. Indeed, the two approaches can be quite similar. This point is illustrated by the close relationship between the ECPR and the Ramsey approach indicated in our formulas (31) and (32). Thus the difference between the two approaches simply has to do with the social cost of public funds. There would be a second source of difference related to demand cross-elasticities, and so on, if the ECPR were identified with the simple margin rule, but to do so would be to use the wrong measure of opportunity cost.

It is clear that for both ECPR-style pricing and Ramsey-style pricing there is a formidable need for information about demand and supply elasticities. Estimation of the relevant elasticities will inevitably be imperfect, and estimation errors imply efficiency losses. The extent of those losses can, however, be diminished if the incumbent's profit on retail sales can be reduced—for example, by allowing cost-reflective tariff rebalancing.

3 Competitive Bottlenecks

This section is a bridge between the two main sections of the chapter, and discusses the case of what might be termed "competitive bottlenecks". The discussion is framed in terms of two applications: call termination on mobile telephony (in section 3.1) and access pricing for competing internet networks (section 3.2). The basic underlying features of these kinds of markets include: (i) there are several networks competing vigorously for the same pool of subscribers, and (ii) even though networks compete vigorously *for* subscribers, they often have a monopoly position in providing communications services *to* their subscribers. In such markets it is undesirable to leave network access arrangements entirely to the discretion of individual networks: there is, at the least, a role for coordination between networks, and in many cases a role for regulatory intervention to set access charges at the appropriate level.

3.1 Mobile Call Termination

First consider the market for mobile telephony.[56] Here the features to emphasize are that each subscriber has only one channel of communication, so that there is no "dual sourcing" of mobile telephony, and that tariff arrangements are such that the caller pays for the whole cost of the call. Given that these conditions are satisfied then, when a subscriber signs up with a network, that network has a monopoly over delivering calls to the subscriber, and it can extract monopoly profits from the callers to this subscriber. Even if the market for subscribers is intense, so that overall profits are eliminated in the sector, these monopoly profits—and the consequent deadweight losses—persist and can be used to finance subsidized retail tariffs offered to attract subscribers.

[56] This section is based on Armstrong (1997). See also Hausman (2000), Wright (1999a) and Gans and King (2000) for further analysis.

This topic is analyzed first in a simple model that makes the main points most clearly, and then various generalizations are discussed. Suppose, then, that there are two telecommunications sectors: fixed and mobile. In this benchmark model the latter is assumed to be competitive, and all mobile charges except possibly for call termination are unregulated. The simplifying assumptions in this section are:

Assumption 1: All calls made from mobile networks are terminated on the fixed sector;
Assumption 2: Mobile subscribers gain no utility from receiving calls;
Assumption 3: Mobile subscribers do not care about the welfare of the people who call them;
Assumption 4: Mobile subscribers do not pay anything for receiving calls made to them;
Assumption 5: The mobile sector is perfectly competitive.

We make Assumption 1 so that the charge for call termination on mobile networks does not affect the cost for making calls from mobile networks.[57]

The cost structure of a mobile network is assumed to consist of a fixed cost k per subscriber, a constant marginal cost c^O for outbound calls (by Assumption 1, these are always made to the fixed sector) and a constant marginal cost c^T for terminating calls from the fixed sector. (All mobile networks are assumed to be identical.) The outbound cost c^O includes any termination payments made to the fixed sector, which are assumed to be constant in this analysis. The fixed per-subscriber cost k is the cost of a mobile handset together with any other costs associated with managing subscribers (such as billing costs). In sum, if a subscriber receives Q calls and makes q calls, a network's costs for that subscriber are $c^T Q + c^O q + k$.

Fixed network operators are required to pay a charge a per call to a mobile network for call termination. Suppose that with this charge the retail price for calls from the fixed network to the mobile network is $P(a)$. (If different mobile networks choose different termination charges, then this is reflected in different call charges from the fixed sector.) With a given fixed-to-mobile price P, suppose each subscriber on a mobile network receives $Q(P)$ calls from the fixed sector. Suppose a mobile network offers its subscribers a fixed charge f and a per-call charge p for making calls. Suppose that once a subscriber has joined a mobile network with per-call charge p, that subscriber makes a quantity $q(p)$ of outbound calls. (This is assumed for now not to depend on the price of incoming calls P.) Then a mobile

[57] If Assumption 1 does not hold then the analysis becomes closely related to the next section on "two way" access pricing. In fact, the final model in section 4.2.4 below is closely related to the model presented here, but allows for traffic between "mobile" networks. The analysis there is simplified by the assumption that the fraction of calls that are made to other subscribers on the *same* mobile network is assumed to be negligible. Wright (1999a) and Gans and King (2000) also allow for mobile-to-mobile calls, and moreover allow subscribers on the same network to call each other. This feature complicates the analysis considerably. However, many of the conclusions of their analysis are similar to those presented here; in particular, they find that profits from high termination charges are used to attract mobile subscribers.

network's profit per subscriber is

$$\Pi = \underbrace{(p - c^O)q(p) + f - k}_{\text{profit from subscription}} + \underbrace{(a - c^T)Q(P(a))}_{\text{profit from termination}} . \tag{41}$$

From Assumption 5 equilibrium in the mobile sector is such that (i) operators' overall profits Π are driven down to zero, and (ii) subscriber utility is maximized subject to a network's break-even constraint. The consumer surplus of each mobile subscriber is $v(p) - f$, where $v'(p) \equiv -q(p)$. From Assumptions 2, 3 and 4 this competitive equilibrium involves marginal cost pricing for outbound calls, so that $p = c^O$, and the fixed subscriber charge recovers any profit shortfall, so that from (41) we have

$$f = k - (a - c^T)Q(P(a)) . \tag{42}$$

This implies that the choice of termination charge a has no effect on the charge for outbound calls p, but only on the fixed charge f. Thus, if a is above the cost of call termination then $(a - c^T)Q(P(a)) > 0$ and a mobile operator subsidizes the retail charge for network connection in the sense that $f < k$. Thus in this particular model handset subsidies, which are a feature of many mobile markets, are a direct result of charging for call termination above cost.[58]

3.1.1 Unregulated Termination Charges

Suppose first that mobile networks are free to choose all their charges, including termination charges. In this model it is clear what will happen: competition will drive overall profits to zero, but networks in equilibrium must maximize their profits from call termination in (41). Therefore, a is chosen to maximize

$$(a - c^T)Q(P(a)) . \tag{43}$$

Call this unregulated termination charge a^{mon} for future reference. Thus, even with perfect competition *for* mobile subscribers, there is no competition for providing access *to* mobile subscribers. In particular, call termination is not charged for at the correct rate, and there is a role for regulatory intervention.[59] The socially optimal access charge is analyzed in the next section.

[58] In some ways this mechanism is similar to that which functions in markets with "switching costs"—see Klemperer (1995). When consumers incur switching costs for changing their supplier for a given service, a supplier has a degree of market power once a consumer has been signed up. Therefore, in competitive markets, suppliers will offer subsidized initial deals, these being funded out of the future profits generated once consumers become locked in. The difference between the two cases is with "switching costs" it is original consumer's "future self" who is exploited, whereas in the "competitive bottleneck" case it is *other* people who suffer.

[59] In some situations this monopoly access price could be extremely high. For instance, Gans and King (2000) discuss the case where the price of calls to mobiles is a function of the *average* cost of call termination across all mobile networks. (One reason for this might be that fixed subscribers are ignorant about the

Before that, however, it is convenient to discuss the case where subscribers *do* care about receiving calls, since this is sometimes used as an argument for why call termination is not necessarily a "bottleneck" in the sector. So suppose that Assumption 2 does not hold, and that each mobile subscriber gains utility b per call from the fixed sector, in addition to the direct utility from making calls of $v(p) - f$. Therefore, with the termination charge a total subscriber utility is $v(p)+bQ(P(a))-f$. Maximizing this expression subject to profits in (41) being non-negative implies that the equilibrium unregulated termination charge maximizes

$$(a - [c^T - b])Q(P(a)) \; . \tag{44}$$

Thus, comparing (43) with (44) we see that the effect of introducing call externalities is indeed to reduce the equilibrium unregulated termination charge. (In fact, the effect is precisely as if the cost of termination is reduced by the factor b.) In the next section we discuss, among other things, whether this is a good argument for not regulating such charges.

3.1.2 Regulated Termination Charges

For now, return to the case where subscribers do not value incoming calls. To derive the optimal termination charge we initially make three further simplifying assumptions:

Assumption 6: The number of mobile subscribers is not affected by tariffs in the mobile market (over the relevant range of tariffs);
Assumption 7: The price of calls from the fixed to the mobile sector is equal to perceived marginal cost;
Assumption 8: Changes in the fixed-to-mobile call price have no effect on the demand for any other services offered by the fixed sector.

Assumption 7 might hold either because of competition in the fixed sector or because of optimal regulation in the fixed sector. If the marginal cost on the fixed network for making calls to the point of interconnect with mobile networks is C, then the total perceived marginal cost of calling a mobile subscriber is $C+a$. Assumption 7 therefore implies that $P(a) \equiv C+a$.

Consumer surplus per mobile subscriber in the market for calls *to* mobile subscribers, when that price is P, is $V(P)$, where $V'(P) \equiv -Q(P)$. Therefore, from (42) and Assumptions 6, 7 and 8, total welfare per mobile subscriber when the termination charge is a is

$$W = \underbrace{V(C+a)}_{\text{utility of callers to mobiles}} + \underbrace{v(c^O) + (a - c^T)Q(C+a) - k}_{\text{utility of mobile subscribers}} \; . \tag{45}$$

mobile network they are trying to call and about the associated tariff. They therefore base their calling decisions on the average charges for calls to mobiles.) In this case a small mobile network's access charge has a negligible effect on the average call termination charge, with the result that its monopoly access charge will be very high indeed.

(Profit in both sectors is zero, and Assumption 8 implies that the fixed-to-mobile price has no effect on other profits in the fixed sector. Assumption 6 implies that welfare per mobile subscriber is an accurate and well-defined measure of total welfare.) This expression is maximized by setting $a = c^T$, so that there should be marginal cost pricing of call termination under our assumptions. This in turn implies that $f = k$ in equilibrium, and there are no handset or other subsidies for mobile network connection at the optimum. Although mobile subscribers certainly benefit from high termination charges—since their network connection is subsidized as a result—this benefit is more than outweighed by the costs this imposes on their callers.

Although this simple model captures some important features of the market, it ignores a number of important aspects. The next section adapts the model to discuss the determination of access charges in more complex environments.

3.1.3 Extensions

Price of fixed-to-mobile calls not equal to cost: Suppose that Assumption 7 does not hold, and that the price of calls to mobile subscribers is above the perceived marginal cost. This might be, for instance, because regulation in the fixed sector required that call charges be above cost in order to fund social obligations, or because this call charge is unregulated. Then, with no other modifications to the above benchmark model, it is optimal to have the termination charge be such that the price of calls to mobile subscribers equal to the *actual* cost of making such calls, i.e. so that $P(a) = C + c^T$. When $P(a) > C + a$ this obviously implies that $a < c^T$, and so it is optimal to subsidize call termination on mobile networks in order to counteract the price-cost mark-up in the fixed-to-mobile call market.

Allowing for network externalities: Suppose next that Assumption 6 does not hold, so that the number of mobile subscribers is elastic.[60] Specifically, if $U = v(p) - f$ is the net surplus offered by the mobile networks, suppose that the total consumer surplus of the mobile subscribers is $\Phi(U)$ and the number of mobile subscribers is $N(U) = \Phi'(U)$, where $N(\cdot)$ is an increasing function. (We assume that potential mobile subscribers differ only in their willingness to subscribe, and not in the number of calls they make once they do subscribe.)

Callers on the fixed network benefit from higher mobile subscription, as there are then more people to call on the mobile network. As above, suppose that $Q(P)$ is the number of fixed-to-mobile calls per mobile subscriber, which is assumed not to depend on the number of mobile subscribers. The fact that all potential mobile subscribers are equally valuable to fixed subscribers, which is what this constant calls per mobile subscriber assumption

[60] Section 5 of Wright (1999a) also discusses this issue, but in a more complicated model where mobile networks have some market power. However, he too obtains the result that when market participation is sensitive to mobile tariffs it is optimal to set the mobile termination charge above cost. More generally, much of the analysis in Willig (1979) is to do with tariff design in the presence of network and call externalities.

implies, means that the "network externality" is linear in the number of mobile subscribers. In particular, consumer surplus in the fixed sector is $NV(P)$ when there are N mobile subscribers and the fixed-to-mobile call charge is P.

As before, overall mobile network profits are zero and mobile subscriber utility is maximized given this zero profit constraint. This again implies that $p = c^O$ and f is given by (42), so that $U(a) = v(c^O) + (a - c^T)Q(C + a) - k$. For values of a less than a^{mon} in section 3.1.1, $U(a)$ is increasing in a since a higher termination charge increases the profits from call termination, which then translates into a lower fixed charge f. This in turn translates into higher subscriber numbers for the mobile sector. Total welfare, which was previously given by (45), is now

$$W = \underbrace{N(U(a))V(C+a)}_{\text{utility of callers to mobiles}} + \underbrace{\Phi(U(a))}_{\text{utility of mobile subscribers}} .$$

Maximizing this with respect to a gives the first-order condition

$$a = c^T + \frac{N'(U(a))U'(a)V(C+a)}{-NQ'(C+a)} \geq c^T .$$

Therefore, unless $N' = 0$, so that mobile subscription is totally inelastic as in the previous section, the network externality effect implies that it is optimal to set the mobile call termination charge *above* cost. The reason for this is clear: a higher termination charge raises the equilibrium mobile subscriber utility via handset subsidies and the like, this in turn increases mobile subscription, which in turn raises the utility of fixed network subscribers because of the network externality effect.

However, even though the presence of network externalities gives a reason for pricing call termination above cost, the optimal termination charge is still lower than the unregulated charge a^{mon} in section 3.1.1 above.[61] Therefore, the realistic assumption that network externalities are important in mobile telephony does not provide a good argument for deregulating mobile call termination charges.[62]

Allowing for call externalities: As was earlier mentioned in section 3.1.1, suppose a subscriber gains utility b from receiving each call. Given Assumption 7 welfare per mobile subscriber is modified from (45) to

$$W = V(C+a) + v(c^O) + bQ(C+a) + (a - c^T)Q(C+a) - k .$$

[61] The regulator would not choose a termination charge above a^{mon} since both mobile subscribers and callers to mobiles would be made better off by reducing a down to a^{mon}. Also, welfare is increased by reducing a at least a little below a^{mon} since this has only a second-order effect on U (and hence on mobile subscriber numbers) and yet yields a first-order benefit for callers to mobile subscribers.

[62] In any event, it is possible that there exist superior ways to fund the (desirable) subsidy to mobile subscribers instead of imposing a price-cost margin on the price of calls from fixed-to-mobile networks. (Perhaps a small tax on the whole industry would be less distortionary than a big tax on one sector?)

The optimal termination charge therefore satisfies the first-order condition

$$a = c^T - b < c^T \ . \tag{46}$$

Therefore, if mobile subscribers derive a benefit from incoming calls, then the regulator should set the termination charge *below* cost in order to encourage calls from the fixed sector.

This result means that the presence of call externalities is *not* a good reason to de-regulate call termination. In section 3.1.1 we showed that call externalities did cause networks to reduce their termination charges, in order to stimulate demand for calls to their subscribers. However, we have also seen that call externalities act to reduce the socially optimal charge. Comparing (46) with (44) we see that unregulated networks still price termination in excess of the socially optimal level.

Allowing subscribers to care about their callers: This extension provides a better argument for de-regulating call termination. Suppose next that Assumption 3 does not hold, and for simplicity suppose that mobile subscribers internalize the entire welfare of those who call them when they choose their mobile network. (For instance, if the regular callers are family members or close friends, this might be a good approximation.) Mobile networks will still compete away all profits, and total welfare of mobile subscribers—which now includes that of callers on the fixed network—is maximized subject to this break-even constraint. Suppose that mobile firms are free to choose their own termination charge. If Assumption 7 continues to hold, then it is clear that marginal cost pricing will be the equilibrium outcome: $p = c^O$, $f = k$ and $a = c^T$. Therefore, if mobile subscribers and those who call them are mostly closely-knit groups, the need for the control of mobile call termination in a competitive market is much reduced.

Allowing for charging for incoming calls: In some countries, such as the United States, the arrangement is that calls from the fixed to the mobile sector are charged at regular fixed network tariffs, and the recipient of the call also makes a payment per call to cover the additional cost of terminating on mobile networks.[63]

To be concrete, suppose that it is now the mobile subscriber who pays for call termination, not the caller. Then Assumption 7 states that with this arrangement we have $P = C$. If a mobile network charges a per call to *its* subscribers for receiving incoming calls, then its overall profit from its subscribers is modified from (41) to be

$$\Pi = (p - c^O)q(p) + f - k + (a - c^T)Q(C) \ .$$

[63]This arrangement is often due to country-specific numbering issues, such as callers being unable to distinguish mobile and fixed telephone numbers due to the former having no distinct number ranges. See Kim and Lim (2001), Hermalin and Katz (2001) and Jeon, Laffont, and Tirole (2004) for analysis of the general issue of reception charges in telecommunications.

If mobile subscribers have no choice but to accept incoming calls, being charged for these calls is exactly equivalent to increasing the fixed charge.[64] (Recall that all mobile subscribers receive the same, known, number of calls in this model.) Therefore, the balance between the fixed charge f and the incoming call charge a is not determined here. Equilibrium here involves $p = c^O$, as usual, and $f + (a - c^T)Q(C) = k$, so that all other charges cover the mobile network's fixed costs.

Clearly the distortion is no longer that mobile operators exploit their market power over delivering calls to their subscribers—that problem has been eliminated in competitive markets by making mobile subscribers themselves pay for call termination—but that callers on the fixed networks now pay too little. (The price P should, in the absence of call externalities, be equal to the total cost $C + c^T$ rather than just C.) Of course, it is just possible that one market failure will cancel out another, and if there is a call externality b that is precisely equal to the termination cost c^T, this receiver-pays tariff arrangement might be desirable. However, the two effects are again quite different, and the presence of call externalities is not in itself an argument for making call recipients rather than callers cover the cost of call termination.

Allowing for partial substitution between fixed and mobile sectors: Finally, suppose that Assumption 8 does not hold. There are several ways in which charges for calls to and from the mobile sector affect the take-up of other services provided by the fixed sector. For instance, low charges for mobile service could cause some subscribers to leave the fixed network altogether and make all their calls on their mobile network, which of course reduces the demand for fixed services. Alternatively, consider a person who subscribes to both fixed and mobile services. Since calls to this person using the two networks are substitutes, we imagine a reduction in the price of fixed-to-mobile calls would, all else equal, reduce the demand for fixed-to-fixed calls. For simplicity, we focus on the second of these two kinds of substitutability.

Suppose that, if the fixed-to-mobile call charge is P, total profits (per mobile subscriber) in the rest of the fixed sector are $\pi(P)$. (We suppose that charges for all other fixed services are not affected by policy towards the mobile sector.) Assumption 7 implies that total welfare is modified from (45) to

$$\pi(C + a) + V(C + a) + v(c^O) + (a - c^T)Q(C + a) - k \; .$$

(Here, $V(P)$ is consumer surplus in the fixed sector keeping all other fixed sector charges constant.) Maximizing this with respect to a implies that

$$a = c^T + \frac{\pi'(C + a)}{-Q'(C + a)} \; . \tag{47}$$

[64] The papers mentioned in the previous footnote allow the call recipient to cut short, or not accept, a call if the charge outweighs the benefit of being called.

Thus the sign of $(a - c^T)$ is the same as that of π'. It is plausible that π' is positive since increasing the fixed-to-mobile price will increase the demand for fixed-to-fixed services, which will usually add to profits. If this is so then it is optimal to set the termination charge above cost, in order to stimulate demand for (profitable) fixed-to-fixed services. Expression (47) is a variant of the ECPR formula (17). In order to provide the marginal unit of access to the fixed sector, the mobile termination charge must increase by $1/Q'$. Therefore, the second term in the above expression is the lost profit in the *fixed* sector caused by the *mobile* sector supplying a unit of access.

3.2 Access Charges for the Internet

A market that in several ways is quite similar to that of the above model of mobile telephony is the internet. A simple model of the internet has two classes of agent: web-site providers (who provide information and content of various kinds) and consumers (who wish to obtain content provided on the web-sites).[65] In this simple model, all information flows are "one way", from web-sites to consumers. Consumers obtain utility from viewing content on the websites, and web-site providers obtain utility from consumers visiting their web-sites. There are several different means by which web-site providers might obtain utility from consumers. A "commercial" web-site might sell content to visitors, either electronically (such as an economics journal being published electronically) or acting as an on-line retailer for other kinds of products (such as *Amazon.com*). Alternatively, a web-site provider might gain utility even if there is no direct payment from the visitor. For instance, it might make money from providing advertisements to its visitors, from providing useful information about a firm's products (which are then purchased by conventional means), from saving money on conventional postage when content is downloaded from the site, or it might simply obtain utility from knowing people have visited the site. For simplicity, in this section we consider the second kind of web-site, and suppose that web-site providers do not charge consumers directly for entry to the site.

Consider first a benchmark model where the number of consumers and the number of web-sites is exogenously fixed (provided that some reservation level of utility is obtained by the two groups), with the number in each group being normalized to 1. Suppose each consumer obtains utility u from visiting each web-site, and each web-site obtains utility $\hat{u}$ from each consumer who visits it. There are a number of identical internet networks operating in a perfectly competitive market. The total cost of carrying a unit of communication from a web-site to a consumer is $c^O + c^T$, where c^O is the cost of *originating* communication from the web-site and c^T is the cost of *terminating* the communication to the consumer. If a consumer connected to one network visits a web-site hosted on a rival network, the host network incurs the cost c^O while the terminating network incurs the cost c^T.

[65] The following discussion is based on Laffont, Marcus, Rey, and Tirole (2001). See also chapter 7 of Laffont and Tirole (2000) and Little and Wright (2000).

Since the act of a consumer visiting a web-site benefits both parties—it is similar to the above model of mobile telephony with the extension to allow for call externalities—it is not obvious whether the terminating network should be paid by the originating network or *vice versa*. Here we adopt the convention that the web-site's network pays the access charge a to the consumer's network for delivering the communication. If a is negative, however, this arrangement implies that the web-site's network is paid for providing the service of delivering the communication to the consumer's network. Turning to the retail side, suppose network i charges its consumers p_i each time they access any web-site and charges its web-sites $\hat{p}_i$ each time any consumer visits them.

Putting all of this together implies that if network i attracts n_i consumers and $\hat{n}_i$ web-sites, its total profits are

$$\Pi_i = \underbrace{\hat{n}_i(\hat{p}_i - c^O - n_i c^T - (1-n_i)a)}_{\text{profits from web-sites}} + \underbrace{n_i(1-\hat{n}_i)(a-c^T)}_{\text{profits from call termination}} + \underbrace{n_i p_i}_{\text{profits from reception charges}} .$$

(Note that the cost allocation involved in this decomposition is quite arbitrary, as we have loaded all of the costs involved in "completing calls" onto the web-site segment of market.) This can be simplified to

$$\Pi_i = n_i(p_i + a - c^T) + \hat{n}_i(\hat{p}_i - a - c^O) . \tag{48}$$

Therefore, the profits of a network can be decomposed into those generated by selling services to web-site providers and those generated by its services to consumers. The effective marginal cost of providing services to consumers is $c^T - a$, while the effective marginal cost of providing services to web-sites is $c^O + a$.

Given the assumption of perfect competition between networks, equilibrium prices are driven down to the associated marginal costs:

$$p_i = c^T - a \; ; \; \hat{p}_i = c^O + a \tag{49}$$

and network profits are zero.[66] Thus, the choice of regulated access charge affects the balance of retail charges offered to consumers and to web-site providers in equilibrium. If there is no access charge for interconnection, so a "bill and keep" system is used, then consumers are charged the cost of terminating communications to them, while web-sites are charged the cost of originating communication. By contrast, if access is charged at termination cost, so that $a = c^T$, then consumers pay nothing for using the internet while web-sites pay the full cost of providing communication. Alternatively, if the originating network can recover its costs, so that $a = -c^O$, then the reverse holds.

[66] For this result to be valid, these candidate prices must not exceed the agents' reservation utilities, so that the access charge should satisfy $\hat{u} - c^O > a > c^T - u$.

Notice in particular that a network's charging strategy involves setting its retail prices as though all terminating traffic came from rival networks and as though all originating traffic was to be terminated on rival networks. For instance, a web-site connected to a given network will have a fraction of visits from consumers connected to the same network, and yet the origination charge is $c^O + a$ which is the cost of sending communications to a rival network. Laffont, Marcus, Rey, and Tirole (2001) call this result the "off-net-cost pricing principle".[67]

While this simple model with inelastic demand is suitable for demonstrating how the access charge feeds through into retail charges in equilibrium, it is incapable of analyzing the normative issue of the optimal level of the access charge. (As long as all consumers and web-site providers are served, welfare is not affected by the balance between the two retail charges.) Therefore, we next extend the model to allow for some responses to prices. One way to do this is to suppose that there is elastic consumer demand for visiting web-sites (but the number of web-sites remains fixed). If the price for receiving traffic is p, suppose that each consumer makes $q(p)$ visits to each web-site. Otherwise, everything is as described above. Then, after some manipulation, (48) becomes

$$\Pi_i = n_i q(p_i)(p_i + a - c^T) + \hat{n}_i \left[n_i q(p_i) + (1 - n_i)q(\bar{p})\right](\hat{p}_i - a - c^O) \ ,$$

where $\bar{p}$ is the (average) retail price for receiving communications for consumers on the rival networks. Unlike the previous case with inelastic consumer demand, here a network's profits do not neatly decompose into profits from consumers and profits from web-sites. Nevertheless, the off-net-cost pricing principle still holds, and equilibrium prices are given by (49) above.[68] Equilibrium profits are zero as before. However, unlike the inelastic case where the level of the access charge had no effect on welfare, here there is a welfare effect. Assuming everyone is served, welfare per web-site with the reception charge p is

$$v(p) + [p + \hat{u} - c^O - c^T]q(p) \ .$$

(Recall that a web-site gains utility $\hat{u}$ from each visit. Here $v(p)$ is the consumer surplus function associated with $q(p)$, so that $v'(p) = -q(p)$.) Maximizing this implies that the ideal reception price is

$$p^* = c^O + c^T - \hat{u} \ ,$$

[67] This insight implies that the market works just as if there are two kinds of network, one kind just caters for consumers and one kind just caters for web-sites. There is one-way traffic from the latter set of networks to the former. Viewed from this perspective, this market is very similar to the model of mobile telephony in the previous section, with call traffic from the fixed networks to the mobile networks. (The fact that mobile networks also called the fixed network played no role in the analysis, since the termination charge on the fixed networks, and hence the quantity of calls to the fixed network, was exogenously fixed.)

[68] For instance, if the rival networks offer a reception price $\bar{p}$ which is greater than the effective marginal cost $c^T - a$, then it pays network i to undercut this price slightly and to take the entire population of consumers.

and so from (49) this implies that the optimal access charge that implements this outcome is

$$a = \hat{u} - c^O \ . \tag{50}$$

This access charge also implements the origination charge $\hat{p} = \hat{u}$, which means that web-site providers are left with none of the gains from trade in equilibrium. This is to be expected: Ramsey principles suggest that any service inelastically supplied should bear the burden of price-cost markups. If the supply of web-sites was also elastic, then the optimal access charge will have to trade-off the efficiency losses of both sides of the market.[69] However, the formula (50) does not necessarily imply that the access charge is high, or even positive: if $\hat{u}$ is small, so that most of the benefits of communication are on the consumer side, then the access charge will be negative, i.e. the originating network is paid for supplying valuable information to consumers.[70]

This brief discussion of some issues concerning access pricing between internet networks is similar in some ways to the earlier analysis of the mobile telephony market. In particular, even though there is effective competition for both consumers and web-site providers, this does not justify a *laissez-faire* policy towards interconnection arrangements between networks. Perhaps the main difference between the mobile model and the internet model is that in the latter there are *two* bottlenecks: (i) once a consumer has signed up with a network, that network has a monopoly over providing communication to that consumer (similar to the case of mobile subscribers above), and (ii) once a web-site has signed up with a network, that network has a monopoly for originating communication from that web-site.[71]

4 Two-way Access Pricing and Network Interconnection

In this section of the chapter we discuss the important topic of two-way network interconnection. In contrast to the scenarios outlined in previous sections, here all firms in the

[69] See Laffont, Marcus, Rey, and Tirole (2001) for this analysis, and also for several further extensions to this basic model (including allowing for imperfect competition between networks, and for differential charging according to whether traffic is "on-net" or "off-net").

[70] As well as this Ramsey analysis, Laffont, Marcus, Rey, and Tirole (2001) discuss how networks would cooperatively choose the access charge a to maximize profits. (They extend this competitive model to allow for imperfect competition.) They show that there is no reason to expect that networks will agree to charge the socially optimal charge, and so there is some scope for regulatory intervention in this market.

[71] This latter effect is also present in the mobile model—once a subscriber has joined a particular mobile network, that network must originate all calls from the subscriber—but is unimportant when there is effective competition. The difference is that in the mobile model people on the fixed network were assumed not to obtain any benefit from being called by mobile subscribers and, more importantly, they were not charged for receiving calls.

market must negotiate with each other to gain access to each other's subscribers.[72] Because this analysis can quickly become complicated, we look first at the case of fixed, captive subscriber bases, which is naturally illustrated by the case of international call termination. In subsequent sections we allow these market shares to be determined endogenously by the competitive process.

4.1 Fixed Subscriber Bases: International Call Termination

Consider two countries, A and B.[73] Suppose that the cost of originating a call in country i (to be terminated in the other country) is c_i^O and the cost of terminating a call in country i (originating in the other country) is c_i^T. Let the total cost of making a call from i to j be $c_i = c_i^O + c_j^T$. The price of a call from country i to country j is p_i. The demand for calls from i to j is $x_i(p_i)$.[74] Let π_i be defined by $\pi_i(p_i) = (p_i - c_i)x_i(p_i)$, and this the profit function in country i if call termination happens to be priced at marginal cost. Consumer surplus in i is $v_i(p_i)$, where $v_i' = -x_i$. (We assume that the prices in this international market do not affect the demand for other telecommunications services supplied in the two countries.) Total (world) surplus due to the call traffic between the countries is therefore

$$v_A(p_A) + \pi_A(p_A) + v_B(p_B) + \pi_B(p_B) \ ,$$

which is maximized by setting prices equal to the actual marginal costs: $p_i = c_i$. Let the call termination charge in country i be a_i. Then the profits of country i due to this international market are

$$\Pi_i = \underbrace{(p_i - c_i^O - a_j)x_i(p_i)}_{\text{profits from call origination}} + \underbrace{(a_i - c_i^T)x_j(p_j)}_{\text{profits from call termination}} \quad . \tag{51}$$

Assume that the move order is for termination charges a_i to be chosen first and then, taking these as given, countries choose their retail prices p_i non-cooperatively.[75] Therefore,

[72] In fact the previous model of internet interconnection also had this feature, in that networks typically would have both consumers and web-sites as customers, and so networks would need access to each others' consumers to deliver communications originating on their own networks. However, in that model originators and recipients of communications were disjoint groups, and information flows were always in one direction. This feature greatly simplifies the analysis.

[73] See Carter and Wright (1994), Hakim and Lu (1993), Cave and Donnelly (1996), Yun, Choi, and Ahn (1997), section 6 of Laffont, Rey, and Tirole (1998a), Box 5.1 in Laffont and Tirole (2000), Domon and Kazuharu (1999), and Wright (1999b) for more analysis of the economics of international settlements in telecommunications. At a deeper level, this analysis is closely related to the problem of negotiating trade tariffs/subsidies between two large countries—for instance, see Mayer (1981) for a classic treatment.

[74] For simplicity, assume there are no cross-price effects for calls in the two directions. See Acton and Vogelsang (1992) for evidence that cross-price effects are not significant.

[75] The reverse ordering (or assuming all prices are chosen simultaneously) does not make sense since if the two retail prices are fixed then the quantity of calls which country i must terminate, which is $x_j(p_j)$, is also fixed. (we assume a country cannot refuse to terminate calls at the specified price.) Country i can then set an arbitrarily high termination charge a_i and make arbitrarily high profits, and no equilibrium can exist with the alternative move order.

given the pair of access charges (a_A, a_B), each country i chooses its retail price p_i to maximise its welfare, taking the other's retail price as given.

Suppose first that retail price regulation (or competition) in each country is such that, given the foreign country's termination charge, the call charge is equal to the perceived marginal cost of the call:

$$p_i = c_i^O + a_j \ . \tag{52}$$

(This is similar to Assumption 7 in section 3.1.2.) Clearly, if

$$a_i = c_i^T \text{ for } i = A, B \tag{53}$$

then both countries will set the ideal prices $p_i = c_i$. Therefore, cost-based termination charges induce the best outcome from the point of view of overall welfare, at least provided national regulators act in the way described above. Given the one-to-one relationship between a_i and p_j in (52) we can think of each country as choosing the *other* country's retail price for calls. Written in this way profits in (51) become

$$\Pi_i = (p_j - c_j)x_j(p_j) = \pi_j(p_j) \ .$$

(There are no profits from call origination.) Welfare in country i is therefore

$$w_i = v_i(p_i) + \pi_j(p_j) \tag{54}$$

where country i chooses p_j and country j chooses p_i.

So far the analysis has been done assuming that regulation or competition in each country forces the price of international calls to be equal to the perceived cost of making such calls. International calls have historically been priced substantially above the associated costs—even taking into account the existing (high) international call termination payments—and so it is worthwhile to extend this analysis to allow for the possibility that countries might wish to use profits from international calls for other, perhaps socially useful, purposes. One way to model this is to suppose that country i receives benefits of $1 + \lambda_i > 1$ for each unit of profit it makes in the international sector, from both retail and call termination sources. (The previous analysis assumed that $\lambda_i = 0$.) In this case welfare in country i is given by

$$w_i = v_i(p_i) + (1 + \lambda_i)\left\{(p_i - c_i^O - a_j)x_i(p_i) + (a_i - c_i^T)x_j(p_j)\right\} . \tag{55}$$

Since profits are now more valuable, a country will wish to set its outbound price above its perceived marginal cost. Indeed, maximizing the above expression with respect to p_i given a_j implies that marginal cost pricing is replaced by the Ramsey formula

$$\frac{p_i - c_i^O - a_j}{p_i} = \frac{\lambda_i}{1 + \lambda_i}\frac{1}{\eta_i} \ , \tag{56}$$

where η_i is the elasticity of demand for calls from i to j. In this case we have $p_i > c_i^O + a_j$, although there is still a one-to-one, increasing relationship between a_j and p_i. A country with

a higher "social cost of public funds" parameter λ_i and/or a less elastic demand for calls will choose a higher price/cost markup for its international calls. If we write

$$\phi_i(a_j) = \max_{p_i} : \left\{ v_i(p_i) + (1+\lambda_i)(p_i - c_i^O - a_j)x_i(p_i) \right\}$$

to be the maximum welfare in country i due to outbound calls, then the envelope theorem implies that

$$\phi_i'(a_j) = -(1+\lambda_i)\hat{x}_i(a_j) \tag{57}$$

where $\hat{x}_i(a_j)$ is the optimal number of calls given the overseas country's termination charge a_j. Here, $\hat{x}_i$ is decreasing in a_j. Also, a higher value for λ_i translates into a higher value for $(p_i - c_i^O - a_j)x_i(p_i)$, which in turn induces a smaller value for $\hat{x}_i$.

4.1.1 Non-cooperative Determination of Termination Charges

Suppose now that each country chooses its termination charge—or equivalently, the overseas country's retail price—non-cooperatively. Given the pricing rule (52), country i knows that if it chooses the overseas retail price p_j it will make a profit from call termination equal to $\pi_j(p_j)$. Since the choice of termination charge does not affect its consumer surplus from originating calls (which is chosen by j), i will choose p_j to maximize profits from call termination, so that p_j is chosen by i to satisfy the usual monopoly formula

$$\frac{p_j - c_j}{p_j} = \frac{1}{\eta_j} > 0\,. \tag{58}$$

This implies that $a_i > c_i^T$.[76] Therefore, if countries choose their termination charges independently, each will set a charge above cost, with the result that world surplus is not maximized. Each country exploits its monopoly position in call termination with the result that, despite the first-best pricing behaviour at the national level in (52), retail prices are set as if networks were unregulated profit-maximizing monopolies in each direction. In particular, it is straightforward to see that each country will benefit if both termination charges are brought down at least a little from this non-cooperative equilibrium, since each country suffers only a second-order loss in profits from call termination, but a first-order gain in the surplus from making calls.[77] Thus we have the important insight that non-cooperative setting of termination charges will cause the charges to be set at too high a level, due to the standard double-marginalization problem.

[76] More generally, if the overseas country has demand $\hat{x}_j(a_i)$ given the home country's termination charge—as illustrated by the case of Ramsey pricing mentioned above—then i will choose a_i to maximize its call termination profit $(a_i - c_i^T)\hat{x}_j(a_i)$.

[77] However, it is *not* true that it is in each country's interest individually to reduce both termination charges all the way down to cost. For instance, suppose one country has only a tiny demand for calls to the other country, so that $v_i \approx 0$. This country will then lose out if both charges are brought down to cost.

4.1.2 Cooperative Determination of Termination Charges

Given the inefficiency of choosing termination charges non-cooperatively, it is natural and desirable that countries negotiate their mutual termination charges. Here we consider three kinds of negotiations with progressively less complex kinds of bargaining:

Bargaining with side-payments: If side-payments can costlessly be made between countries (and if there is no asymmetric information about costs and demands), it is natural to suppose that an efficient outcome is attainable, and termination charges will be set equal to costs: $p_i = c_i$ and $a_i = c_i^T$. How the first-best surplus is divided between the countries will depend on the details of the bargaining procedure.

Bargaining with non-reciprocal termination charges: Next, suppose that no such side-payments are possible, and the two countries simply bargain over the pair of access charges (assuming retail prices are then set as in (52)). Whatever the precise form the negotiations take, it is reasonable to suppose bargaining is (second-best) efficient in the sense that one country's welfare is maximized subject to the other's welfare being held constant, i.e. retail prices p_A and p_B must maximize w_A subject to $w_B \geq \bar{w}_B$ for some reservation level $\bar{w}_B$. From (54) the first-order conditions for this problem are

$$\frac{\pi'_A(p_A)}{x_A(p_A)} = \frac{x_B(p_B)}{\pi'_B(p_B)}$$

or

$$\left(1 - \frac{p_A - c_A}{p_A}\eta_A\right)\left(1 - \frac{p_B - c_B}{p_B}\eta_B\right) = 1 \ .$$

In particular, either $p_i = c_i$ for both countries (which can be optimal only in knife-edge situations, depending on $\bar{w}_B$) or one country's call charge is above cost and the other's is below cost. Therefore, except in totally symmetric situations (when $p_i = c_i$ and $a_i = c_i^T$ is the outcome of any reasonable bargaining process), one country makes a loss on terminating calls, and the other makes a profit.

Bargaining with reciprocal termination charges: Finally, suppose that countries must choose symmetric termination charges, so that $a_A = a_B = a$, say. Given the pricing rule (52), welfare in (54) with a reciprocal termination charge a is

$$w_i = v_i(c_i^O + a) + \pi_j(c_j^O + a) \ ,$$

and so country i's ideal *reciprocal* termination charge, denoted a_i^*, is given by the expression

$$a_i^* = c_i^T + \frac{x_j(c_j^O + a_i^*) - x_i(c_i^O + a_i^*)}{-x_i'(c_j^O + a_i^*)} \ . \tag{59}$$

Therefore, in symmetric situations where $c_A^T = c_B^T$, $c_A^O = c_B^O$ and $x_A \equiv x_B$ the interests of countries coincide, and each is happy to agree to charge for call termination at marginal cost. In other cases, however, countries will have divergent interests. Given divergent preferences, it is natural to suppose that the equilibrium reciprocal charge will lie between the two privately preferred values, a_A^* and a_B^*, and its precise location will depend on the balance of "bargaining power" between the two countries.

Expression (59) shows that a country would like to have a reciprocal termination charge that is higher than its termination cost whenever (i) originating call costs are approximately equal and the foreign country has higher demand for calls than the home country, or (ii) demand functions are approximately equal and the foreign country has lower costs for originating calls than the home country. In sum, countries with either high costs or a net inflow of calls over the relevant range of prices will prefer a higher reciprocal termination charge. In practice this means that poorer countries will, on the whole, prefer higher reciprocal termination charges than will more developed countries.

More generally, when the countries have a social cost of public funds, welfare (55) in country i with the reciprocal charge a is

$$w_i = \phi_i(a) + (1 + \lambda_i)(a - c_i^T)\hat{x}_j(a) \ .$$

From (57), this is maximized by setting

$$a_i^* = c_i^T + \frac{\hat{x}_j(a_i^*) - \hat{x}_i(a_i^*)}{-\hat{x}_i'(a_i^*)} \ ,$$

which generalizes (59) above. This shows that another factor which causes preferences to diverge over termination charges is the social cost of public funds. As discussed above, in otherwise symmetric environments, if $\lambda_j > \lambda_i$ then $\hat{x}_j(a) < \hat{x}_i(a)$ and so country i prefers a lower reciprocal termination charge than country j. Since poorer countries will also tend to have a higher social cost of public funds—due, for instance, to fewer sources for effective taxation—this will be yet another reason to expect that these countries will prefer higher reciprocal termination charges.

4.2 Interconnection with Competition for Subscribers

In this section we extend the analysis to allow networks to compete for subscribers, rather than taking market shares as exogenously fixed. While some of the insights from the analysis of international call termination continue to be valid in this competitive case—for instance, non-cooperative setting of termination charges will lead to inefficiently high retail prices—others will not.

4.2.1 A General Framework

Consider the following model: there are two networks, A and B, competing for the same subscribers.[78] Each subscriber purchases all telephony services from one network or the other (or none). There is a continuum of potential subscribers, with total number normalized to 1. Subscribers are differentiated, and if they receive "utility" u_A from using network A and u_B from using network B, then network A has $n_A = s_A(u_A, u_B)$ subscribers and network B has $n_B = s_B(u_B, u_A)$ subscribers. (These utilities will be derived below.) Naturally, $s_i(u_i, u_j)$ is increasing in u_i and decreasing in u_j, since a better deal being offered by the rival causes a network's subscriber numbers to fall. If $n_A + n_B = 1$, then all potential subscribers join one or other network over the relevant range of utilities; otherwise there is only partial participation, and network externalities become an important ingredient of the analysis.

An example of the market share function s_i is obtained from the familiar Hotelling model of consumer choice. Suppose that the two firms are "located" at each end of the unit interval. A consumer's type (or location) is denoted $y \in [0, 1]$. If the two firms' utilities are u_A and u_B, then such a consumer gains utility of $u_A - wy$ if she joins network A, and utility $u_B - w(1-y)$ if she joins network B. Here $w > 0$ is a parameter that determines how closely substitutable are the two services. Suppose consumers' types are uniformly distributed over the unit interval. Then s_i is given by

$$n_i = s_i(u_i, u_j) = \frac{1}{2} + \frac{u_i - u_j}{2w} \tag{60}$$

(provided that $0 \leq n_i \leq 1$). This is an example where there is full subscriber participation, so that $n_A + n_B \equiv 1$ over the relevant range of utilities.[79]

Depending on the context, networks could be using two-part pricing or linear pricing.[80] To keep the analysis as general as possible at this stage, suppose firm i offers its subscribers the tariff

$$T_i(x, \hat{x}) = p_i x + \hat{p}_i \hat{x} + f_i \ , \tag{61}$$

where x is the number of calls made to other subscribers on the same network i ("on-net calls"), $\hat{x}$ is the number of calls made to subscribers on the rival network ("off-net calls"), p_i is the marginal price for on-net calls, $\hat{p}_i$ is the marginal price for off-net calls, and f_i

[78] This basic model is adapted from Armstrong (1998), Laffont, Rey, and Tirole (1998a) and Laffont, Rey, and Tirole (1998b). The first paper considered only linear retail prices, whereas the latter two analyzed the more relevant case of two-part tariffs as well. The last of these discussed the case where networks are permitted to condition their call charges upon the destination network. See also the closely related work of Carter and Wright (1999), who examine in more detail whether firms might not interconnect at all. See Chapter 5 in Laffont and Tirole (2000) for another overview of two-way network interconnection issues. The "competition in utility space" approach used in this section is explored in more detail in Armstrong and Vickers (2001).

[79] More generally, one can show that whenever there is full subscriber participation, subscriber decisions depend only upon the *difference* in utilities $u_i - u_j$.

[80] The case of fully nonlinear pricing is considered below in section 4.2.3.

is the fixed charge (which is zero if linear pricing is used). Public policy (or technological limitations) may require that $p_i = \hat{p}_i$, so that networks cannot price discriminate according to the destination network. We assume that subscribers do not pay for receiving calls.[81]

A subscriber is assumed to gain the same utility from calling every other subscriber, and if a subscriber faces the price p for calling some other subscriber, the former will make $x(p)$ calls to the latter. This demand function is assumed to be independent of the identities of the caller and recipient, and also of the network that each party has joined. (Later we will allow callers to differ in their demand for calls and in how many calls they receive.) Let $v(p)$ be the level of consumer surplus associated with the demand function $x(p)$, so that $v' \equiv -x$. For now, suppose that subscribers obtain no utility from receiving calls—this assumption is relaxed in the next section. Then, given the subscription choices made by the other subscribers, the utility received by a subscriber if she joins network $i = A, B$ is

$$u_i = n_i v(p_i) + n_j v(\hat{p}_i) - f_i \ . \tag{62}$$

Thus a subscriber's utility is linear in the number of subscribers of each network, i.e. network externalities take the same linear form as we used in the mobile telephony discussion in section 3.1.3. Notice that even if there is full subscriber participation, so that $n_A + n_B = 1$, whenever $p_i \neq \hat{p}_i$ there are what Laffont, Rey, and Tirole (1998b) term "tariff mediated network externalities" present, and subscribers will choose their network partly on the basis of the number of other subscribers on the network. The system of choices in (62) is consistent provided that

$$n_A = s_A(u_A, u_B) \ ; \ n_B = s_B(u_B, u_A) \ .$$

Therefore, given the pair of tariffs offered by the firms, this system of four equations in four unknowns will, at least in the cases we consider, yield the unique equilibrium subscriber numbers. For instance, in the simple Hotelling model (60) one obtains[82]

$$n_A = 1 - n_B = \frac{m_A - \frac{1}{2w}(f_A - f_B)}{m_A + m_B} \tag{63}$$

as the unique equilibrium, where

$$m_i = \frac{1}{2} + \frac{v(\hat{p}_i) - v(p_j)}{2w} \ .$$

(Here we require that $0 \leq n_A \leq 1$.)

The *net* number of calls from network i to network j, which we denote by z_i, is

$$z_i \equiv n_i n_j (x(\hat{p}_i) - x(\hat{p}_j)) \ . \tag{64}$$

[81] See Kim and Lim (2001), Hermalin and Katz (2001) and Jeon, Laffont, and Tirole (2004) for analyses of the case where networks can charge for incoming calls.

[82] See Section 3 of Laffont, Rey, and Tirole (1998b) for more details, including a discussion of the important issue of whether this equilibrium is stable.

The function z_i represents what might be considered the "net demand for access" by network i. Note that when networks choose the same call charges—so that in particular $\hat{p}_i = \hat{p}_j$—then the net traffic flow between the two networks is zero even if networks have different subscriber numbers.

Turning to the cost side, just as in sections 3.1 and 4.1 above, let c_i^O be network i's cost of supplying a call which originates on its network and which is terminated on the rival network (not including the termination charge), let c_i^T be network i's cost of terminating a call from the other network, and let the cost of originating and terminating a call entirely within network i just be $c_i^O + c_i^T$. (Thus, there is no cost advantage in carrying a call over a single network rather than over two.) Let k_i be the fixed cost of connecting any subscriber to network i. Therefore, as in section 3.1, if a subscriber makes q calls and receives Q calls, the total physical costs for network i are $c_i^T Q + c_i^O q + k_i$.

Let a_i be the charge for terminating a call on network i. Therefore, given the two retail tariffs, the resulting market shares, and the two termination charges, total profits for network i are

$$\Pi_i = \underbrace{n_i \left\{ \left[p_i - c_i^O - c_i^T \right] n_i x(p_i) + \left[\hat{p}_i - c_i^O - a_j \right] n_j x(\hat{p}_i) + f_i - k_i \right\}}_{\text{profit from subscription}} + \underbrace{n_i n_j (a_i - c_i^T) x(\hat{p}_j)}_{\text{profit from termination}} . \tag{65}$$

Similarly to the function π_i in section 4.1, introduce the notation

$$\pi_i(p_i) \equiv (p_i - c_i^O - c_i^T) x(p_i)$$

for the profit function for on-net calls. Then we can rearrange (65) to give

$$\Pi_i = n_i \left\{ n_i \pi_i(p_i) + n_j \pi_i(\hat{p}_i) + f_i - k_i \right\} + n_i n_j \left\{ (a_i - c_i^T) x(\hat{p}_j) - (a_j - c_i^T) x(\hat{p}_i) \right\}$$

which, when termination charges are reciprocal (i.e. when $a_A = a_B = a$), simplifies to

$$\Pi_i = n_i \left\{ n_i \pi_i(p_i) + n_j \pi_i(\hat{p}_i) + f_i - k_i \right\} - (a - c_i^T) z_i . \tag{66}$$

As in section 4.1, the move order is that access charges a_i are chosen first and then, taking these as given, firms choose their retail tariffs non-cooperatively.[83]

[83] A subtle point is that one has to take care about the choice of strategic variables for the firms when network effects are present. For instance, in (62) above there is a one-for-one relationship between utility u_i and the fixed charge f_i. Given $\{p_i, \hat{p}_i\}$ two competitive scenarios are (i) firms offer utilities u_i and then choose f_i ex post in order to deliver the promised utility (which would then depend on the market shares achieved), and (ii) firms offer the fixed charges f_i, and consumers predict the equilibrium market shares and choose their network accordingly. Unfortunately, the outcome is different in the two cases. When this is an issue we will assume that competition takes the form (ii), so that firms compete in tariffs rather than utilities, since that is (perhaps) economically the more plausible.

4.2.2 The First-Best Outcome

Here we consider the benchmark case where the regulator can control the two firms' retail tariffs directly, and so we calculate the socially optimal two-part tariffs, as in (61), for the two firms. From (65), total industry profits with a given pair of retail tariffs are

$$\begin{aligned} &n_A \left\{ \left[p_A - c_A^O - c_A^T \right] n_A x(p_A) + \left[\hat{p}_A - c_A^O - c_B^T \right] n_B x(\hat{p}_A) + f_A - k_A \right\} \\ &+ n_B \left\{ \left[p_B - c_B^O - c_B^T \right] n_B x(p_B) + \left[\hat{p}_B - c_B^O - c_A^T \right] n_A x(\hat{p}_B) + f_B - k_B \right\} \ . \end{aligned}$$

Using the notation $w(p,c) \equiv x(p)(p-c) + v(p)$ and $c_{ij} = c_i^O + c_j^T$ for the cost a making a call from network i to j, and substituting for f_i in (62), this expression for total industry profits can rewritten as

$$\begin{aligned} &n_A \left\{ n_A w(p_A, c_{AA}) + n_B w(\hat{p}_A, c_{AB}) - u_A - k_A \right\} \\ &+ n_B \left\{ n_A w(\hat{p}_B, c_{BA}) + n_B w(p_B, c_{BB}) - u_B - k_B \right\} \end{aligned}$$

Given the two utilities u_A and u_B offered by the firms, let total subscriber surplus be $V(u_A, u_B)$. This function satisfies the usual envelope conditions

$$\frac{\partial}{\partial u_A} V(u_A, u_B) = s_A(u_A, u_B) \ ; \ \frac{\partial}{\partial u_B} V(u_A, u_B) = s_B(u_A, u_B) \ .$$

Total welfare is therefore

$$\begin{aligned} W = V(u_A, u_B) &+ n_A \left\{ n_A w(p_A, c_{AA}) + n_B w(\hat{p}_A, c_{AB}) - u_A - k_A \right\} \\ &+ n_B \left\{ n_A w(\hat{p}_B, c_{BA}) + n_B w(p_B, c_{BB}) - u_B - k_B \right\} \ . \end{aligned}$$

Clearly, for any pair of utilities $\{u_A, u_B\}$, this expression is maximized by choosing each firm's call charges to maximize the relevant welfare function $w(\cdot, c_{ij})$, i.e., to equal the relevant marginal cost:

$$p_i = c_i^O + c_i^T \ ; \ \hat{p}_i = c_i^O + c_j^T \ . \tag{67}$$

In particular, when termination costs differ on the two networks, we see that price discrimination according to the destination network is socially optimal.

With the prices in (67) total welfare becomes

$$\begin{aligned} W = V(u_A, u_B) &+ n_A \left\{ n_A v(c_{AA}) + n_B v(c_{AB}) - u_A - k_A \right\} \\ &+ n_B \left\{ n_A v(c_{BA}) + n_B v(c_{BB}) - u_B - k_B \right\} \ . \end{aligned}$$

Maximizing the above with respect to u_i, and making the substitution in (62), one obtains the following expressions for the optimal fixed charges:[84]

$$f_A = k_A - n_A v(c_{AA}) - n_B v(c_{BA}) \ ; \ f_A = k_A - n_A v(c_{AB}) - n_B v(c_{BB}) \ . \tag{68}$$

[84] Since there is no strategic interaction between the firms in this welfare analysis, the issue mentioned in footnote 84 does not arise, and maximizing welfare with respect to fixed charges and with respect to utilities yields the same result.

(Of course, the subscriber numbers n_A and n_B are endogenous in the above, and are determined jointly with f_A and f_B.)

In sum, (67) shows that call charges should equal the relevant marginal costs, while (68) shows that the fixed charge should be subsidized below the fixed cost to reflect the externality that a subscriber's choice of network exerts on other callers. (If a subscriber chooses to join network A, say, then each of A's other subscribers obtains benefit $v(c_{AA})$ and each of B's subscribers obtains benefit $v(c_{BA})$, which therefore yields the required subsidy in (68).) In particular, we see that—except in a special case discussed below—it is not optimal to have marginal cost pricing for all services, and the charge for joining a network should differ from cost in order, roughly speaking, to attract more subscribers onto the network with the lower *termination* cost.

The formula (68) is valid for all network choice functions $s_i(u_A, u_B)$, and in general this first-best policy calls for subsidies to be provided out of public funds. However, in the important special case where total subscriber numbers are fixed, so that $n_A + n_B \equiv 1$, then policy is simplified. In particular, since all subscribers opt to join one or other network, market shares depend only on the difference in utilities. Therefore, adding a constant amount to both networks' fixed charge in (68) will not affect subscriber decisions, or total welfare, and so the first best can be achieved without recourse to subsidies.

This special case of full participation also makes clear that it is the difference in the firms' termination, rather than origination or fixed, costs that is the motive for the distortion in (68). For instance, suppose that the firms have the same termination cost, and so in particular that $c_{iA} \equiv c_{iB}$. Then the first-best is achieved by pricing all services, including the fixed charge of joining a network, at marginal cost. In particular, in the symmetric case when the two firms have the same costs—which is the focus of the next section—the optimal outcome is indeed implemented by marginal-cost pricing. The reason for this is that, when call charges are given by (67), differences in origination costs are fully "internalized" by subscribers at the time they choose their network.

By contrast, when firms differ only in their termination costs, then with a marginal-cost pricing regime, subscribers do not fully take into account the implications of their network choice on their callers. (This situation is somewhat related to the model of the mobile market in section 3.1 above.) To correct for this, the optimal fixed charge/fixed cost margin is lower on the network with the lower termination cost, as in (68).

A natural question to ask is whether this first-best outcome can be implemented by means of a suitable choice of termination charges, and then leaving the two networks to compete at the retail level. Unfortunately, the answer to this, except in a few special cases, is *No*. (In the next section, where the focus is on symmetric situations, we will see that the first-best can often be implemented when the termination charge is equal to cost.) The reason is that the access charges are being call upon to perform too many tasks. The first best involves the correct choice of six variables (the two pairs of call charges for the two firms together with the two fixed charges), and yet we have only two instruments (the pair of termination charges) at our disposal. In fact, under a cost-based termination charge regime, where $a_i \equiv c_i^T$, then

the four call charges will be at the first-best level, given in (67), in equilibrium. However, there is no reason to believe that the fixed charges that emerge from competition will be equal to (68).

No price discrimination: The above analysis assumed that firms used the most general tariffs as in (61). Since this form of price discrimination involving different call charges for on-net and off-net calls is not common at present, it is worthwhile to perform this welfare analysis under the assumption that firms must charge the same for all calls (i.e., that $p_i \equiv \hat{p}_i$). For simplicity, suppose that there is full subscriber participation, so that $n_A + n_B \equiv 1$. Then one can show that (67) becomes

$$p_A = n_A c_{AA} + n_B c_{AB} \; ; \; p_B = n_A c_{BA} + n_B c_{BB} \; .$$

Therefore, the call charge on network i is set equal to the *average* marginal cost as weighted by market shares, so that $p_i = n_A c_{iA} + n_B c_{iB}$. And, corresponding to (68), when call charges must be uniform we see that the optimal pattern of fixed charge/fixed cost markups is given by

$$\begin{aligned} (f_A - k_A) - (f_B - k_B) &= n_A x_A \left[c_{AA} - c_{AB}\right] + n_B x_B [c_{BA} - c_{BB}] \\ &= [n_A x_A + n_B x_B][c_A^T - c_B^T] \end{aligned} \tag{69}$$

where we have written $x_i = x(n_A c_{iA} + n_B c_{iB})$ for the equilibrium number of calls made by a subscriber on network i.

If the right-hand side of (69) is negative, then subscriber access to network A is subsidized relative to network B. Clearly, this is optimal if and only if call termination on network A is lower than on B. The intuition for this is precisely the same as for the previous case in (68) where networks offered more flexible tariffs: with marginal cost pricing, subscribers do not adequately take account of the effect their choice of network has on their callers' call charges. Therefore, incentives to join a network are distorted away from marginal costs in order to induce more people to join the network with the lower termination cost, as shown in (69).

4.2.3 Symmetric Competition: A Danger of Collusion?

Can networks that compete for subscribers be relied upon to choose termination charges that are close to socially optimal? Or can networks use the choice of termination charges to affect their equilibrium behaviour in the retail market to boost profits at the expense of welfare? We will see in the following succession of models that the answers to these questions are rather subtle, and depend (i) on whether networks use linear rather than nonlinear tariffs and (ii) on whether they price discriminate between on-net and off-net calls.

In this section we suppose that the two networks are symmetric in terms of cost, so that $c_A^O = c_B^O = c^O$, $c_A^T = c_B^T = c^T$ and $k_A = k_B = k$. Write

$$\pi(p) \equiv x(p)(p - c^O - c^T) \tag{70}$$

for the on-net profit function for each network. We also use the (symmetric) Hotelling formulation for subscriber choices as in (60). Because of symmetry we assume that networks agree to charge a reciprocal access charge a to each other.[85]

Linear non-discriminatory pricing: Here, despite its lack of realism, suppose that the networks are constrained to offer only linear tariffs, so that $f_i = 0$. We also suppose that networks are prohibited from engaging in price discrimination according to the destination network, so that $p_i \equiv \hat{p}_i$. In this case, since the total number of subscribers is always equal to 1, (66) simplifies to

$$\Pi_i = n_i \{\pi(p_i) - k\} - (a - c^T) z_i \ . \tag{71}$$

Since total industry profits are $\pi(p_A)+\pi(p_B)$, the joint-profit maximizing (or collusive) linear retail price, denoted p^*, is the price that maximizes π in (70).

Consider the sub-game in which firms choose retail tariffs given that a reciprocal access charge a has initially been chosen. From (71), the first-order condition for p to be the equilibrium choice for each firm is that $\partial \Pi_i(p,p)/\partial p_i = 0$. With the Hotelling specification (60), this entails

$$-\frac{x(p)}{2w}\left(\pi(p) - k\right) + \frac{1}{2}\pi'(p) - \frac{1}{4}(a - c^T)x'(p) = 0 \ . \tag{72}$$

Since joint profits are maximized at $p = p^*$, where $\pi'(p^*) = 0$, if collusion *can* be sustained by choosing a suitable access charge, say a^*, from (72) this access charge must be

$$a^* = c^T + \frac{x(p^*)}{-x'(p^*)}\frac{2}{w}\left(\pi(p^*) - k\right) > c^T \ . \tag{73}$$

In particular, we see that $a^* > c^T$. This candidate for the collusive termination charge is high when (i) $\frac{1}{w}$, which is a measure of the substitutability of the two networks' services, is high, (ii) when demand is inelastic, i.e. when $-x/x'$ is high, or (iii) when $\pi(p^*) - k$, the maximum profit per subscriber, is high. Note that in the limit as w becomes very large, so that market shares are fixed, the collusive termination charge is equal to cost. This is an instance of the result in section 4.1 that in symmetric situations the ideal reciprocal termination is equal to cost—see (59) above.

When the access charge is set according to the rule (73), firms have no *local* incentive to deviate from the collusive price p^* in the retail market, even though retail prices are set non-cooperatively: if one firms undercuts the other by a small amount, the gain in retail profits from increased market share is just offset by the increased access payments needed for the increased number of calls made to the rival network.

[85] It is intuitive that if networks set access charges non-cooperatively, they will set them *higher* than if they negotiate over a common charge, since there will be a serious "double marginalization" problem. This is exactly comparable to the case of international call termination discussed in section 4.1.1 above. See also Laffont, Rey, and Tirole (1998a) for more details. It will therefore be socially desirable to allow firms to "collude" over the choice of the access charge compared to the case where each firm acts independently.

The remaining question is when the first-order condition (72) does in fact characterize the (globally) optimal response for one firm given that the other has chosen p^*. It is easy to see that services being close substitutes will make this collusion impossible. From (60), if one network deviates and chooses the price $p_L < p^*$, it can take the whole market provided that $v(p_L) \geq v(p^*) + w$. Suppose also that this price p_L is fairly close to p^* in the sense that $\pi(p_L) > \frac{1}{2}\pi(p^*)$. If such a price may be found—and clearly this is possible if w is sufficiently small—then the collusive price p^* cannot be sustained in equilibrium with any access charge, for it would always pay a firm to corner the market by choosing the price p_L (which thereby eliminates the need for access payments altogether).[86] On the other hand, if products are not close substitutes then the sacrifice in retail profits needed to come close to capturing all subscribers is too great and no undercutting of the collusive price p^* is unilaterally profitable.

Therefore, if the services offered by the two networks are sufficiently differentiated, setting a high termination charge may indeed be used as an instrument of collusion. The reason for this is that high termination charges increase the cost of reducing retail prices unilaterally, since such an action causes the deviating network to have a net outflow of calls to the rival network, which then results in costly call termination payments. In sum, a major contrast between the one-way and two-way models of access pricing is that in the former the chief danger is that high access charges can be used as an instrument of *foreclosure*—perhaps to drive out or otherwise disadvantage rivals in the downstream market—whereas in the latter case high access charges can sometimes be used as an instrument of *collusion*.[87] An implication is that firms should not be free to set their termination charges, even in the case where there is no dispute between firms.

In fact this result is not really surprising. It is very plausible that different termination charges will affect the choice of retail tariffs offered by competing networks in equilibrium. For instance, when price discrimination is banned—so that $\hat{p}_i = p_i$—the choice of a directly affects the average cost of a call, and this will naturally feed through into a higher equilibrium call charge. What is more surprising is that this result—that access charges can implement collusion—is easily overturned in natural extensions to this model, as will be seen in the next sections.

Two-part tariffs without network-based price discrimination: Next consider the more realistic

[86] More precisely, if a departs significantly from termination cost then equilibrium in the retail market simply does not exist when the market is sufficiently competitive—see Laffont, Rey, and Tirole (1998a) and Laffont, Rey, and Tirole (1998b) for several instances of this kind of problem. An interesting feature of the recent analysis in Laffont, Marcus, Rey, and Tirole (2001), which was discussed in section 3.2 above, and Jeon, Laffont, and Tirole (2004) is that, even in perfectly competitive markets, equilibrium exists when the access charge departs from cost. In these two latter papers networks charge for receiving calls, whereas in the earlier papers this price was missing. Introducing this price makes competition more "stable"—see the discussion in section 2 of Laffont, Marcus, Rey, and Tirole (2001) for details.

[87] This kind of analysis can also be applied to the pay-TV market—see section 2 of Armstrong (1999) for a model in which two TV companies agree to pay each other large amounts for each other's programmes in order to weaken competition for subscribers.

case where networks compete using two-part tariffs. (We maintain the assumption that price discrimination according to destination network is banned.) We will show that the collusive impact of the access charge is now eliminated, at least in the simple symmetric and full-participation model we are using. Indeed, it is straightforward to show that whenever a symmetric retail equilibrium exists, the resulting profits cannot depend at all on the choice of termination charge.

With a given reciprocal termination charge a, network i's profit in (71) is modified to be

$$\Pi_i = n_i \{\pi(p_i) + f_i - k\} - (a - c^T) z_i \ . \tag{74}$$

Suppose that the symmetric retail equilibrium involves each network offering the two-part tariff $T(x) = px + f$, in which case the equilibrium subscriber utility is $u_A = u_B = v(p) - f$. Consider what happens if one firm decides to vary its fixed charge f a little, keeping the call charge constant at p. (Clearly at any equilibrium, a firm should have no incentive to deviate from f.) Because each network offers the same call charge p, the net demand for access z_i in (64) is zero, even when firms have different market shares. Therefore, firms can compete for market share—by reducing f—*without* incurring call termination payments, and this is the major difference with the linear pricing case.

It is now straightforward to derive equilibrium profits. With the market share specification (60), when a firm offers the fixed charge $\hat{f}$ rather than the equilibrium charge f, its total profit in (74) is

$$\left(\frac{1}{2} + \frac{f - \hat{f}}{2w} \right) \left(\pi(p) + \hat{f} - k \right) \ .$$

For f to be an equilibrium this expression must be maximized at $\hat{f} = f$, and so we obtain the first-order condition

$$\pi(p) + f - k \equiv w \ . \tag{75}$$

Since the left-hand side of this expression is the total industry profits (which is split equally between the two networks), we see that the choice of termination charge a has no effect on profits. In particular, a cannot be used as a collusive device when non-discriminatory two-part tariffs are used.

The choice of a *does* affect the equilibrium choice of the price p, however. Indeed, Proposition 7 in Laffont, Rey, and Tirole (1998a) shows (using the current notation) that the equilibrium call charge is

$$p = c^O + \frac{1}{2}(c^T + a) \ , \tag{76}$$

which is just the perceived average cost of a call when networks divide the market equally.[88] Therefore, just as in the above linear pricing case, a high access charge feeds through into a high retail call charge, and hence leads to high profitability from *calls*. However, the

[88] This proposition also shows that equilibrium exists when w is not too small or $a - c^T$ is not too large.

additional instrument of the fixed charge is then used to attract these profitable subscribers, with the result that these excess profits are partially competed away.[89] Clearly, in this model the socially optimal retail price is $p = c^O + c^T$, and so the socially optimal access charge is equal to cost: $a = c^T$. Since the firms are indifferent between termination charges when two-part tariffs are used, in theory they will not object to a regulatory suggestion to set $a = c^T$. In this sense, then, competition for subscribers greatly simplifies the regulatory problem when two-part tariff are used.

Two-part tariffs with call externalities: To see how robust the above "profit-neutrality" result is, consider the following extension. Suppose that each subscriber obtains utility b for each call received. As above, networks charge a two-part tariff without price discrimination according to destination network. In this case subscriber utility in (62) is modified to

$$u_i = v(p_i) - f_i + b\left[n_i x(p_i) + n_j x(p_j)\right] \ .$$

Here, the term in $[\cdot]$ is the total number of calls received by a subscriber, and this is *independent* of the chosen network. Thus, a subscriber's utility from receiving calls, which does in general depend on the market shares of the two networks, does *not* bias a subscriber's network choice at all.[90] Since with the Hotelling specification in (60) market shares depend only on the difference in offered utility levels, the presence of call externalities has no effect on the competitive outcome. In particular, the equilibrium profits and retail tariffs are still given by (75) and (76), and are not affected by the externality parameter b.

What is affected, however, is the socially optimal termination charge. With the call externality b, the socially optimal call charge is reduced to $p = c^O + c^T - b$ in order to stimulate call volume. From (76) this implies that the socially optimal termination charge is below cost:[91]

$$a = c^T - 2b \ .$$

Again, though, the fact that profits are unaffected by the termination charge suggests that firms should not object to this regulatory requirement to subsidize network interconnection.[92]

[89] This effect is similar to that seen in the analysis of the mobile market in section 3.1. There, networks make large profits from terminating calls which are then (fully) competed away at the retail level.

[90] In the analysis by Jeon, Laffont, and Tirole (2004) of call externalities and the effects of charging for incoming calls, this feature of the market plays a central role. Call externalities represent a *direct* externality, whereas charging for incoming calls creates a *pecuniary* externality. In most cases, those authors show that only the latter are relevant for firms' and subscribers' incentives.

[91] Comparing this reduction to that in the mobile sector—see expression (46)—we see that the subsidy is required to be twice as great. The reason for this is that, in the current context, it is required that a network's perceived (average) marginal cost for a call be reduced by b. However, since only half of a network's calls are destined for the rival network—and so only half incur the cost a—the reduction in a must be correspondingly amplified.

[92] When the analysis is extended to allow for charging for incoming calls, as in Jeon, Laffont, and Tirole (2004), then there are two available instruments for implementing the desired call charge. For instance, one

Nonlinear pricing with heterogeneous subscribers:[93] Another important extension to the basic analysis is to introduce more differences in subscribers. The analysis until now has homogeneity in the calling patterns of subscribers. In particular (i) all subscribers made the same number of calls and (ii) all subscribers received the same number of calls. In this section we extend the analysis to allow for subscriber heterogeneity. Again, though, we do not permit networks to discriminate according to the destination network. For simplicity, we return to the case where subscribers do not care about receiving calls (i.e. $b = 0$).

Specifically, subscribers can differ in their demand for calls, and a *high* (respectively *low*) demand type of subscriber is denoted by H (L). The fraction of subscribers with high demand is α. It is possible that high demand subscribers also differ in the number of calls they receive, and suppose that a fraction β^k of the calls made by type k subscribers are made to type H subscribers, and the fraction $1 - \beta^k$ of their calls are made to low demand consumers. (If $\beta^k \equiv \alpha$ then we would have a model where subscribers were equally likely to call each other subscriber, regardless of the demand characteristics of the call recipient.) Because of this heterogeneity, networks will choose to offer a pair of contracts, one for each of the two kinds of subscriber. Without loss of generality, these contracts specify a total number of calls X in return for a charge f. Suppose that network i offers the type k subscriber a contract allowing X_i^k calls in return for a charge f_i^k. We assume that in equilibrium both demand types of subscriber are served, so that we do not have to consider network externality effects.[94]

We suppose that network choices continue to be made according to the Hotelling model, i.e. that if u_i^k is the *maximum* utility that a type k subscriber can obtain from the contracts offered by network i, then the fraction of type k subscribers who subscribe to network i is given in (60). (Implicitly, we are assuming that the "vertical" differentiation parameter $k = L, H$ is uncorrelated with the "horizontal", or brand preference, parameter y. Therefore, knowledge of brand preference tells a firm nothing about a subscriber's likely demand for calls.) It actually makes no difference to the following argument whether or not networks can observe the demand type of any given subscriber.[95] However, if a subscriber's demand characteristics *are* private information then networks must ensure that the incentive constraints are satisfied, and that a type k subscriber finds it privately optimal to choose the contract (X_i^k, f_i^k) instead of the contract aimed at the other demand type.

---

could have call termination at cost, and impose a suitable incoming call charge. When, however, there is the possibility that some subscribers will refuse to accept incoming calls rather than pay the incoming call charge, then it is superior not to charge for incoming calls.

[93] This is adapted from Dessein (2003) and Hahn (2004). See also section 3.5 of Tirole (1988) for an account of the approach to nonlinear pricing used here.

[94] See Poletti and Wright (2003) for analysis of the case where participation constraints have an impact on networks' policies. They show that the profit neutrality result fails in this case.

[95] Or, in the terminology often used in the literature, it makes no difference whether the model is one of third degree or second degree price discrimination.

If network i has a fraction n_i^k of the demand type k subscribers, the total number of calls received by a type H subscriber (on any network) from callers on network j can be shown to be

$$n_j^H \beta^H X_j^H + \frac{1-\alpha}{\alpha} n_j^L \beta^L X_j^L \ ,$$

while the number received by a type L subscriber (on any network) from callers on network j is

$$\frac{\alpha}{1-\alpha} n_j^H (1-\beta^H) X_j^H + n_j^L (1-\beta^L) X_j^L \ .$$

The number of calls made to network j by a type k subscriber on network i is

$$X_i^k \left[n_j^H \beta^k + n_j^L (1-\beta^k) \right] \ .$$

Therefore, with these call volumes and market shares, network i's net outflow of calls—i.e. its demand for access z_i—is given by the complex expression

$$\begin{aligned} z_i = {} & \alpha n_i^H X_i^H \left[n_j^H \beta^H + n_j^L (1-\beta^H) \right] + (1-\alpha) n_i^L X_i^L \left[n_j^H \beta^L + n_j^L (1-\beta^L) \right] \\ & -(1-\alpha) n_i^L \left[\frac{\alpha}{1-\alpha} n_j^H (1-\beta^H) X_j^H + n_j^L (1-\beta^L) X_j^L \right] \ . \\ & -(1-\alpha) n_i^L \left[\frac{\alpha}{1-\alpha} n_j^H (1-\beta^H) X_j^H + n_j^L (1-\beta^L) X_j^L \right] \ . \end{aligned} \tag{77}$$

If the reciprocal call termination charge is a, then, similarly to (66), network i's profit is

$$\Pi_i = \alpha n_i^H \left[f_i^H - k - X_i^H (c^O + c^T) \right] + (1-\alpha) n_i^L \left[f_i^L - k - X_i^L (c^O + c^T) \right] - (a - c^T) z_i$$

where z_i is as in (77).

Notice that when $X_A^k = X_B^k = X^k$ say, so that the two networks each offer the same pair of quantities in their contracts, the expression for z_i in (77) simplifies to

$$z_i = (n_i^H n_j^L - n_i^L n_j^H)[\alpha X^H (1-\beta^H) - (1-\alpha) X^L \beta^L] \ . \tag{78}$$

In particular, when we also have $n_i^L = n_i^H$, so that network i does not attract a disproportionate number of one demand type of subscriber, then $z_i = 0$.

Next consider equilibrium profits for a given reciprocal access charge a. Given the symmetry between networks, suppose that each network offers the same pair of contracts $\{(X^L, f^L), (X^H, f^H)\}$ in equilibrium. Suppose that network i deviates from this candidate equilibrium by adding an amount ε to both fixed charges f^L and f^H. (In the case where the subscriber demand type k is private information, this uniform change to the fixed charges has no effect on the relative merits of the two contracts, and so does not affect incentive compatibility.) Since we assume that competition causes the participation constraints not to bind for subscribers, this modification does not drive any subscribers from the market, although it

obviously drives marginal subscribers from one network to the other. From (60) this modification means that the network loses an *equal* proportion of both types of subscriber.[96] In particular, from (78) we see that $z_i = 0$ and the deviating firm incurs no access deficit. Similarly to (63), then, profits with this deviation are

$$\Pi_i = \alpha\left(\frac{1}{2} - \frac{\varepsilon}{2w}\right)\left[f^H + \varepsilon - k - X^H(c^O + c^T)\right] + (1-\alpha)\left(\frac{1}{2} - \frac{\varepsilon}{2w}\right)\left[f^L + \varepsilon - k - X^L(c^O + c^T)\right] .$$

For the pair of contracts $\{(X^L, f^L), (X^H, f^H)\}$ to be an equilibrium, it is necessary that this expression be maximized at $\varepsilon = 0$, which just as in (75), yields the first-order condition

$$\alpha\left[f^H - k - X^H(c^O + c^T)\right] + (1-\alpha)\left[f^L - k - X^L(c^O + c^T)\right] = w .$$

The left-hand side of this expression is just the total profits in the market, which is equal to the product differentiation parameter w. Therefore, equilibrium profits again are unaffected by the level of the termination charge a. Thus we see that the profit-neutrality result is not an artifact of the assumption that subscribers were homogeneous.

As in the case of two-part tariffs, the choice of a *does* affect the choice of contracts offered, and $\{(X^H, f^H), (X^L, f^L)\}$ will be a complicated function of the termination charge. The socially optimal choice for a will therefore be the choice that implements the best pattern of consumption. In the case where $a = c^T$, it is possible to show that firms in equilibrium will simply offer the cost-based two-part tariff $T(x) = (c^O + c^T)x + w + k$, so that calls are charged at (actual and perceived) cost, and the firms make a profit of w per subscriber (just as in the previous section on two-part tariffs).[97] Therefore, as before the socially optimal access charge is $a = c^T$, and since firms are indifferent between all levels of the access charge, in principle they will not object to this regulatory policy.

This analysis of two-part tariffs and nonlinear pricing seems to suggest that the choice of termination charge cannot affect profits at all, and networks will not object to a regulatory suggestion to price interconnection at cost (or below cost in the case of call externalities). However, it is important to stress that this convenient result is non-robust in a number of dimensions. For instance the assumed cost and demand symmetry across networks plays an important role in the argument. (Without this it is unlikely that networks will choose reciprocal access charges.) Another reason for non-neutrality is explored in the next section.[98]

[96] Dessein (2003) looks at the case where different subscriber types also have different transport cost parameters w. For instance, high demand subscribers might be more price sensitive. In this case the argument fails, and the access charge has an effect on profits.

[97] See Armstrong and Vickers (2001) and Dessein (2003) for further details.

[98] A further framework in which the profit-neutrality result is unlikely to hold is when there is only partial

Two-part tariffs with network-based price discrimination: Here we return to the homogeneous subscriber framework, but now allow networks to make their call charges depend on the destination network, i.e. to set $p_i \neq \hat{p}_i$. Although this problem looks rather complex, the analysis is simplified by the following observation: in equilibrium all call charges are equal to perceived marginal costs, so that

$$p_A = p_B = c^O + c^T \; ; \; \hat{p}_A = \hat{p}_B = c^O + a \; . \tag{79}$$

(The reasoning for this is the same as that used to derive (67).)

Armed with this observation, expressions (63) and (65) simplify to

$$n_i = \frac{1}{2} + \frac{1}{2}\frac{f_j - f_i}{w + v(c^O + a) - v(c^O + c^T)}$$

and

$$\Pi_i = n_i(f_i - k) + n_i n_j (a - c^T) x(c^O + a) \; .$$

Therefore, the symmetric equilibrium choice of $f_A = f_B = f$ is given by $f = k + w + v(c^O + a) - v(c^O + c^T)$.[99] The resulting total industry profits are then

$$\Pi = w + v(c^O + a) - v(c^O + c^T) + \frac{1}{2}(a - c^T) x(c^O + a) \; . \tag{80}$$

Clearly, this profit does now depend on a, and so the profit-neutrality result does not hold when this form of price discrimination is permitted.

When $a = c^T$ we obtain the same profit w as in (75) for the no-discrimination case analyzed above. (In this case firms do not choose to practice price discrimination, even if they are permitted to do so.) However, networks can do better than this when price discrimination is allowed. For profit in (80) is maximized by choosing the termination charge

$$a = c^T - \frac{x}{-x'} < c^T \; ,$$

and so profits are at their maximum if call termination is *subsidized*.[100]

---

subscriber participation in the market, i.e. when the total number of subscribers rises as the available consumer surplus increases. (This feature was discussed in section 3.1.3 on the mobile sector.) In this case there will be a "market expansion" effect as well as a "business stealing" effect of the choice of access charge, and the former will most likely make equilibrium profits depend on the access charge. The analysis of this case appears to be somewhat technical, and for some preliminary results see Dessein (2003).

[99] We are ignoring the details about when such an equilibrium exists—see Proposition 5 in Laffont, Rey, and Tirole (1998b) for a discussion of this.

[100] There is an error in the proof of Proposition 5 in Laffont, Rey, and Tirole (1998b) which led those authors to conclude that choosing $a = c^T$ maximized profits. This has been corrected in Gans and King (2001). The latter paper argues that the "bill and keep" system for network interconnection—where each network delivers the other's traffic for free—might be a reasonable approximation to this profit-maximizing termination charge (especially if traffic monitoring costs are taken into account).

The analysis presented here assumes that firms used the fixed charge f_i as the strategic variable—see footnote 84 above. Intriguingly, if firms instead used utilities u_i as their strategic variable, then one can show that setting $a = c_T$ becomes the profit-maximizing choice.

Clearly, social welfare is maximized in this model by setting $a = c^T$, and so there is once again a role for network access regulation in this market.[101] Thus, in direct contrast to the linear pricing case discussed above, where high access charges were used to bolster profits at the expense of welfare, here *low* access charges are privately desirable but socially costly. The intuition for this perhaps surprising result is that when $a < c^T$ it is cheaper for subscribers to call people on the rival network than people on their own network, and so, all else equal, subscribers prefer to belong to the *smaller* network. (Recall that, regardless of market shares, each network sets the same pair of call charges given by (79).) This means that the market exhibits "negative network externalities", and firms have little incentive to compete aggressively for subscribers. In effect, by means of a suitable choice of termination charge firms can coordinate on whether to have a market with positive network externalities (high a), no network externalities ($a = c^T$) or negative network externalities (low a). Because of its softening effect on competition, it is mutually profitable for firms to choose the third option.[102]

Discussion: We have seen that the relationship between access charges and equilibrium profits—and in particular whether access charges can be used to sustain "collusion"—depends, in part, on the kinds of tariffs the firms are able to offer. One way to get an intuition about this is to use the framework of "instruments and objectives" in section 2.2.3: when do the firms have enough instruments to attain a given (perhaps collusive) outcome? Since our analysis has assumed the use of only a single instrument—the reciprocal access charge a—firms are likely to be able to sustain a given retail equilibrium only in the case of linear pricing. With linear pricing there is only one objective—the maximization of equilibrium profits within the (restrictive) class of outcomes possible with linear tariffs—and this single instrument will often be enough to implement the optimum. (However, we saw that sometimes equilibrium did not exist for a wide enough range of access charges for this argument to work.)

Other kinds of tariffs, such as two-part tariffs, have at least two dimensions to their definition, and so the single instrument of the access charge will not in general be sufficient to induce a given retail equilibrium. In particular, the joint-profit-maximizing two-part tariff will not in general be sustainable by a particular choice for a.[103] However, it is one

[101] Alternatively, public policy could prohibit this kind of price discrimination, which acts to restore the profit neutrality result, thereby making the regulatory task easier.

[102] Laffont, Rey, and Tirole (1998b) also discuss the case of *linear* pricing with price discrimination. Proposition 2 in that paper shows that when price discrimination is allowed and w is quite large, networks will choose to set the termination charge above cost. Also, Proposition 3 shows that when w is large allowing price discrimination is good for welfare (keeping the termination charge fixed).

[103] If firms did have another instrument at their disposal, then, at least in some circumstances, we would expect collusion to be possible when two-part tariffs are used. For instance, if firms signed a reciprocal agreement so that, say, firm A paid firm B a specified amount for each subscriber A served, then this would be a means of controlling the fixed charge element of the equilibrium two-part tariff. (The access charge then is left to control the usage charge in the usual way.) In a sense, this additional instrument performs a similar

thing to note that the access charge will not be able to sustain the *maximum* profit, and quite another to obtain the result that the access charge has *no effect* on equilibrium profits. This striking profit neutrality result was obtained for the case of both two-part tariffs, both with and without call externalities, and for fully nonlinear pricing. (However, it did not hold for the case of network-based price discrimination). Although there has been little work done, so far, outside the frameworks discussed in this section, one might conjecture that this neutrality result is *very* special, and will depend, for instance, on the specific Hotelling subscriber choice rule we used (involving full subscriber participation). If so, then there will remain a role for regulation of the termination charge.

4.2.4 Asymmetric Competition and Non-Reciprocal Access Charges

The previous section analyzed symmetric competition between networks, and so might be relevant to situations where competition is well-established and mature. In earlier stages of market liberalization, however, competition if left to itself is likely to be skewed in favour of the incumbent, and the analysis needs to be extended to cover such, often more relevant, situations. Unfortunately, there have been only few contributions to the theory of asymmetric competition between networks, so the following discussion is more speculative and incomplete.

Perhaps the central reason why a proper analysis of asymmetric markets is so hard is that there is no reason *a priori* to suppose that access charges should—either from the firms' point of view or in terms of overall welfare—be set reciprocally.[104] Even if the two networks' termination costs (or other costs) were assumed to be the same—itself a questionable assumption if one firm is established and the other is an entrant—this assumption of reciprocity is not innocuous. For instance, in the model with a regulated incumbent firm described below the regulator does not want generally to choose termination charges reciprocally, even when termination costs are the same for the two networks. The discussion of international call termination in section 4.1.2 was also in the context of asymmetric networks, but that framework is a good deal more straightforward than this case where firms compete for subscribers. However, even in that simple framework we saw that the outcome depended on the details of the bargaining process generating the equilibrium. The advantage of symmetry in section 4.2.3 is that the firms' interests coincide at the stage where access charges are determined, and so this bargaining stage is trivial. In asymmetric situations the details of the procedure for choosing access charges will be more important. One of the most valuable

role to the "output tax" instrument discussed in section 2. However, there are a number of reasons—ranging from difficulties involved in monitoring subscriber numbers, to competition law—why such contracts may not be possible.

Another possible instrument is a charge for incoming calls. When this is used, Jeon, Laffont, and Tirole (2004) show that it is sometimes possible for networks to obtain the joint profit-maximizing outcome with a suitable choice of termination charge and reception charge.

[104] Another reason why the asymmetric analysis is not transparent is that marginal cost pricing is then generally not the first-best outcome—see section 4.2.2 above.

areas for future research will be to find compelling theoretical models of how non-reciprocal access charges are determined: ideally tractable models can be found; if not, then numerical simulations seem likely to be most promising way forward.[105]

In the remains of this section we briefly discuss two related models.[106] The first is a model where the two networks choose their retail tariffs without regulatory control, and the second is a model where the dominant firm's retail tariff is controlled.

A simple model of asymmetric competition and non-reciprocal termination charges: The crucial simplification that we make here, and in the next model, in order to make the analysis tractable is to suppose that subscribers have *inelastic* demand for calls.[107] Specifically, suppose that, regardless of the price per call, each subscriber wishes to make exactly one unit of calls to each other subscriber. (We also assume there is full subscriber participation, so that $n_A + n_B \equiv 1$.) In this case, all that matters for a subscriber's network decision is the total charge a network levies for making a single unit of calls to all other subscribers.[108] Suppose, then, that this combined charge is P_i on network i. If $\bar{u}$ is some (high) gross utility a subscriber receives from making his calls, then the utility received by a subscriber if she joins network $i = A, B$ in (62) becomes

$$u_i = \bar{u} - P_i \ . \tag{81}$$

Let a_i be the charge for terminating a call on network i. Therefore, total profits for network i are simplified from (65) to become

$$\Pi_i = \underbrace{n_i \left\{ P_i - \left[c_i^O + c_i^T \right] n_i - \left[c_i^O + a_j \right] n_j - k_i \right\}}_{\text{profit from subscription}} + \underbrace{n_i n_j (a_i - c_i^T)}_{\text{profit from termination}} \ . \tag{82}$$

Introducing the notation

$$C_i \equiv c_i^O + c_i^T + k_i \ ,$$

[105] For some preliminary work in this direction using a dynamic model of network entry, see de Bijl and Peitz (2002) and de Bijl and Peitz (2001).

[106] Other treatments of asymmetric competition are section 7 in Laffont, Rey, and Tirole (1998a) and section 6 in Laffont, Rey, and Tirole (1998b), although the focus there is mainly on the case of reciprocal access pricing. The source of the asymmetry there is the fact that the incumbent has full geographic coverage, whereas the entrant has to install coverage (with a convex cost in so doing). Carter and Wright (2003) present a model of vertical differentiation, in which the incumbent offers a superior service to the entrant. (The cost functions of the two firms do not differ.) Again, though, the analysis focuses on reciprocal access charges. They find that the "larger" firm would like the reciprocal access charge to be equal to the firms' termination cost. Moreover, they show that when the market is very asymmetric, the smaller firm would also like the access charge to be set equal to cost.

[107] This section has particularly benefitted from comments from Julian Wright.

[108] In fact, this is not quite true. If firms charged differently for on-net and off-net calls, then subscribers would have to come to a view about the equilibrium market shares before they can choose their network. Similarly to the discussion in footnote 84 above, this would affect the equilibrium. For simplicity, we sidestep this issue and assume that firms compete in 'total charges' P_i which are invariant to the realized market shares.

and using the assumption that $n_A + n_B \equiv 1$, implies that we can rearrange (82) to give

$$\Pi_i = n_i \{P_i - C_i\} + n_i n_j \{a_i - a_j\} \ . \tag{83}$$

In particular, the outcomes (prices, profits, market shares, welfare) depend only on the *difference* between termination charges on the two networks.

In order to introduce a demand-side asymmetry between the two networks, we can modify the symmetric Hotelling formulation in (60) to give a bias in favour of, say, firm A.[109] Therefore, suppose that if the two firms' utilities are u_A and u_B, the subscriber located at y obtains utility $u_A - ty + t\beta$ if she joins network A, and utility $u_B - t(1-y)$ if she joins network B. If $\beta > 0$ then, with equal utilities offered, firm A will obtain the higher market share. With the two utilities $u_i = \bar{u} - P_i$, the market shares n_i are then given by

$$n_A = \frac{1+\beta}{2} + \frac{P_B - P_A}{2w} \ ; \ n_B = \frac{1-\beta}{2} + \frac{P_A - P_B}{2w} \ . \tag{84}$$

After substituting these market shares into (83), given an initial choice of termination charges $\{a_A, a_B\}$ some tedious calculations show that the equilibrium prices for the two firms are

$$\begin{aligned} P_A &= \frac{2C_A + C_B + \beta w}{3} + w + \frac{\Delta^a}{3}\left\{\beta - \frac{\Delta^C}{w}\right\} \\ P_B &= \frac{C_A + 2C_B - \beta w}{3} + w + \frac{\Delta^a}{3}\left\{\beta - \frac{\Delta^C}{w}\right\} \ . \end{aligned} \tag{85}$$

Here, we have written $\Delta^C = C_A - C_B$ for the cost difference and $\Delta^a = a_A - a_B$ for the difference in termination charges.

Notice that the termination charge differential, Δ^a, affects the two firms' equilibrium prices in the *same* way. This implies that the equilibrium price difference, $P_A - P_B$, is

$$P_A - P_B = \frac{1}{3}\left\{\Delta^C + 2\beta w\right\} \ . \tag{86}$$

Crucially, this price difference, which is what determines the equilibrium market shares, does *not* depend on the termination charges. The equilibrium market shares of the two firms are

$$n_A = \frac{1}{2} + \frac{1}{6}\left\{\beta - \frac{\Delta^C}{w}\right\} \ ; \ n_B = \frac{1}{2} - \frac{1}{6}\left\{\beta - \frac{\Delta^C}{w}\right\} \ . \tag{87}$$

[109] This is the specification for network choice used in Carter and Wright (1999) and Carter and Wright (2003). The following analysis does not depend at all on this particular Hotelling specification, and will apply to any case where there is full participation. We use this functional form merely to derive a closed-form equilibrium.

Without loss of generality, suppose we label the firms so that

$$\beta w > \Delta^C \, . \tag{88}$$

This implies that A's advantage on the demand side—all else equal, subscribers are willing to pay βw more for A's service than for B's—outweighs any cost disadvantage Δ^C. From (87) this implies that A has the larger market share in equilibrium. In addition, for the preceding analysis to be valid we require that the larger firm does not 'corner the market', i.e., that $n_A < 1$ at the equilibrium. Clearly, this requires that the asymmetries are not too great, in the sense that parameters satisfy

$$\beta w - \Delta^C < 3 \, . \tag{89}$$

When (88) holds, (85) implies that *both* retail prices increase as firm A is paid relatively more than B for call termination. In this sense, non-reciprocal access charges, with the larger firm being paid a higher access charge, act as an 'instrument of collusion', in a similar way to the linear pricing model in section 4.2.3. The reason why having large Δ^a acts to relax competition for subscribers is quite intuitive. From (83), when Δ^a is large firm A would like to *increase* the volume of off-net traffic. However, off-net traffic, which is just $n_A n_B$ in this model, is maximized when market shares are equal, and so this gives the larger firm an incentive to make the market more symmetric, i.e., to compete less hard. On the other hard, the smaller firm would like to *decrease* the volume of cross-network traffic as it makes a loss on this traffic when Δ^a is large. This implies that the smaller firm would like to move further from the symmetric allocation, which again gives it an incentive to compete less hard. This is why an increase in Δ^a induces both firms to raise their prices.

However, and in contrast to the symmetric analysis of section 4.2.3, high retail prices are not sufficient to ensure that both firms' profits are high, since we have to take into account the payments for call termination as well. (With non-reciprocal charges, payments for network access do not cancel out in equilibrium, even if off-net traffic is the same in the two directions.) Since termination charges do not affect market shares, it is straightforward to derive the effect on profits for each firm. Since P_A increases with Δ^a, and also, from (83), the profits of the larger firm A are directly increasing in Δ^a, it follows that this larger firm would like the difference Δ^a to be set as large as possible. However, although the retail price of the smaller firm B also increases with Δ^a, from (83) the direct effect of increase Δ^a is negative. To derive the net effect of increasing Δ^a on B's equilibrium profits, note that its profits decrease with Δ^a whenever $P_B - n_A \Delta^a$ decreases with Δ^a (where P_B is an increasing function of Δ^a). However, this is the case whenever (89) holds, which we have assumed. Therefore, the smaller firm would like the difference Δ^a to be set as *low* as possible.

In sum, as one might expect, the two firms have divergent preferences over how termination charges should be set, and regulation of some form is likely to be needed to resolve disputes. (However, since total industry profits do increase with Δ^a, it would be profitable

for the larger firm to compensate by means of a side payment the smaller firm in order to induce the latter to accept a high Δ^a.)

Although total welfare (profits plus subscriber utility) is affected by the market shares of the two networks, for a *given* market share the inelastic demand assumption means that welfare is constant. Therefore, since termination charges cannot be used to alter market shares in this model, total welfare also is unaffected by the choice of termination charges. Thus, in contract to the symmetric situations analyzed in the previous section where a profit-neutrality result was derived, in this asymmetric model we obtain a *welfare*-neutrality result.

However, consumer welfare in isolation is affected by the choice of termination charges, and both retail prices are an increasing function of Δ^a. Therefore, consumer welfare is increased by choosing a lower value for Δ^a. To illustrate this point, suppose that firm A's advantage stems at least in part from the cost side, so that $C_A < C_B$. In this case, consumer welfare is higher with a "cost-based" termination charge regime—so that $a_A < a_B$—than with a reciprocal charging regime ($a_A = a_B$).

A final question to ask is: how does the equilibrium outcome compared with the first best? Using the arguments used in more general settings in section 4.2.2, the first best is achieved when[110]

$$P_A - C_A = P_B - C_B \ . \tag{90}$$

However, (86) implies that

$$P_A - C_A = P_B - C_B + \frac{2}{3}\left\{\beta w - \Delta^C\right\} \ .$$

Therefore, since we have assumed that firms are labelled so that the term $\{\cdot\}$ is positive, we see that competitive outcome involves the larger firm setting too *high* a price compared to its rival than is socially optimal. Put another way, competition results in an outcome that is not "asymmetric enough" from the point of view of overall welfare.

A model with a regulated incumbent: To provide a second example of asymmetric competition between networks, consider the following model with a regulated incumbent facing a fringe of entrants. (Here we have a fringe of entrants, rather than a single rival, in order to abstract from the issue of the market power of the entrant—see footnote 7 above.) As in section 2, the fringe is modeled as consisting of a large number of small networks, each offering exactly the same service and having the same cost functions as each other, and where this service is differentiated from that of the incumbent. In all respects, including the critical assumption of inelastic demand for calls, the model is as presented above. The incumbent firm A is assumed to be regulated at the retail level, and is required to offer the combined charge P_A to its subscribers. This tariff is assumed to be held constant in the following analysis. Its rivals are a competitive fringe, denoted B. This fringe is unregulated at the retail level, and

[110] In fact, this is an instance of the second-best pricing rule in (5) above.

each firm in the fringe offers the combined charge P_B. As before, the utility of a subscriber on the incumbent's network is $u_A = \bar{u} - P_A$, and the utility of those who go to the fringe entrants is $u_B = \bar{u} - P_B$. The market share of the incumbent is n_A, while the fringe takes the remaining $n_A = 1 - n_A$ subscribers.

We look for an access pricing regime that implements the socially optimal outcome, which we know requires that the fringe price satisfies the equal property in (90). Suppose that the termination charges on A and B are a_A and a_B, respectively. Suppose that all firms within the fringe are "small" in the sense that a negligible fraction of calls that originate on a fringe network is terminated on the same network. In that case, analogously to (82), the profit per subscriber for the fringe is

$$\Pi_B = \underbrace{P_B - \left[c_B^O + a_A\right] n_A - \left[c_B^O + a_B\right] n_B - k_B}_{\text{profit from subscription}} + \underbrace{a_B - c_B^T}_{\text{profit from termination}}$$

which can be re-written as

$$\Pi_B = P_B - C_B - \Delta^a n_A \, .$$

Competition within the fringe implies that each firm makes zero profits. This implies that the equilibrium retail charge offered by the fringe is

$$P_B = C_B + \Delta^a n_A \, .$$

Again, this has the same feature as in (85) that B's equilibrium price is an increasing function of Δ^a. However, the reason for this in this case is different: the more a fringe firm receives for terminating calls on its network, the lower its retail charge has to be in order to break even. Therefore, the reason is the same as in the model of the mobile market in section 3.1, where high termination payments were used to fund retail subsidies in equilibrium.

Since, from (90), we wish to implement the fringe price $P_B = C_B + [P_A - C_A]$, we see that the optimal pair of termination charges satisfies

$$a_A - a_B = \frac{P_A - C_A}{n_A} \, .$$

In particular, a reciprocal termination charge is optimal when $P_A = C_A$, i.e., when the incumbent is *optimally* regulated. Perhaps counter-intuitively, then, when there are no regulated distortions at the retail level, it is *not* the case that a cost-based termination regime, where $a_i = c_i^T$, is likely to be optimal.

In other cases, however, non-reciprocal termination charges are used to implement the optimal second best price P_B in (90) given the distorted price charged by the incumbent. If the incumbent is profitable in this market ($P_A > C_A$) then it should charge more for call termination than the fringe, regardless of the underlying relative costs of call termination.

5 Conclusion: Instruments and Objectives

This chapter has had a number of objectives. In particular, relative to most of the existing writing on access pricing, I have aimed to do the following:

1. Pay more attention to the issue of network bypass. When bypass is taken into account, access charges that differ from the incumbent's cost (for instance, ECPR-style access pricing) might have the unfortunate effect of inducing inefficient use (or lack of use) of the incumbent's network.

2. Relatedly, make a more forceful case for pricing access at cost, with the important proviso that suitable retail-level instruments are made available to correct for the incumbent's retail tariff distortions. (Output taxes or subsidies levied on entrants, perhaps in the form of a universal service fund, were suggested for this purpose.)

3. Relate the contentious ECPR policy more clearly to familiar, and well-accepted, principles concerning the theory of the second best.

4. Observe that the more tasks that the access charge is required to perform unassisted—and in the chapter these included (a) the need to give entrants good make-or-buy incentives, (b) the need to induce a desirable amount of entry, given the incumbent's retail tariff distortions, and (c) the need to control the incumbent's retail prices when those were not directly controlled by regulation—the more complicated, and informationally-demanding, the various access pricing formulas become.

5. Present a unified treatment of the two way access pricing problem, encompassing a number of recent theoretical contributions. In particular, it was clearly seen that "easy" policy conclusions—such as (i) access charges can be used to sustain collusive outcomes or (ii) access charges have no effect on equilibrium profits—were not robust to small changes in the assumptions.

6. Argue that, even in the most "competitive" markets—such as the mobile sector—there is likely to remain a role for regulation of access charges.

The instruments used to try to achieve these objectives have been kept as simple and streamlined as possible. Most importantly, I have discussed only the *theory* of access pricing; difficulties involved in implementation have been conveniently ignored. There are numerous further modelling assumptions that have been made, I believe, merely for presentational clarity and tractability. These include:

Full information about costs: I assumed throughout that the incumbent's costs were known to all, and also could not be affected by, say, managerial cost-reducing effort. While it is clear that imperfect regulatory knowledge of costs and the potential for cost reduction has

an important impact on regulatory policy, the *interaction* of these features with the access pricing problem does not often seem to generate many new insights. Thus, it could be said that, to the extent they depend on the realized costs of the incumbent, access pricing regimes such as cost-based access, ECPR or Ramsey pricing all exhibit features of "cost-plus" regulation. In particular, if an incumbent is required to pass on efficiency savings to entrants in the form of lower access charges, it may have poor incentives for cost-reduction. (This is particularly true in the access pricing context, where the incumbent is required to pass on reductions to its *rivals*, rather than merely final consumers as in most monopoly models of optimal regulation.) While this is obviously true, these problems can be tackled using familiar (but imperfect) methods, such as basing access charges on estimated efficient costs, perhaps including computer generated "engineering models" or benchmarking from observed costs in other countries. The global price-cap proposal of Laffont and Tirole is another response to this issue, in that the permitted set of access charges and retail prices do not depend—in the short run, at least—on the firm's realized costs.[111]

Constant marginal costs: Similarly, marginal costs were assumed to be unaffected by the scale of output. This makes little difference to the analysis. For instance, with the competitive fringe model the analysis would go through if the terms C_1 and C_2 were just interpreted as the (endogenous) marginal costs evaluated at the equilibrium.[112]

Other assumptions are far from innocuous, however, and the strong focus on *static* analysis is perhaps the leading limitation of the analysis. For instance, all costs of the incumbent were taken to be avoidable if that firm ceased supplying the relevant service. One could take the convenient view that costs such as C_1 and C_2 are just taken to represent the forward-looking, avoidable component of the firm's costs, and as such are the relevant costs when discussing the efficiency of *future* entry given that important parts of the incumbent's network are already sunk. This, however, is not at all satisfactory. Were the initial investments made with full knowledge of the future access regime? (If so, then all costs are avoidable *ex ante* and the analysis of this chapter is relevant.) This is unlikely, though, given the often long-lived nature of many infrastructure investments in the industry, together with the typical long-run unpredictabilty of regulatory policy. However, to use *ex post* avoidable costs as basis for access pricing policy (to be determined once investments are sunk) is a recipe for opportunism and "deregulatory takings" of the kind emphasized in the writings of Sidak and Spulber. An important next step for research in this area is, I believe, to provide a proper analysis of the *dynamics* of access pricing, focussing on the need to provide long-run, stable incentives for the incumbent (and other firms) to invest efficiently in infrastructure and innovation.

[111] See section 4.7 in Laffont and Tirole (2000) for further discussion. See also Laffont and Tirole (1994) for a full account of optimal access pricing with asymmetric information, and De Fraja (1999) for related analysis.

[112] See Armstrong, Doyle, and Vickers (1996) for analysis along these lines.

References

Acton, J., and I. Vogelsang (1992): "Telephone Demand Over the Atlantic: Evidence From Country-Pair Data," *Journal of Industrial Economics*, 40(3), 305–323.

Armstrong, M. (1997): "Mobile Telephony in the UK," Regulation Initiative Discussion Paper no.15, London Business School.

——— (1998): "Network Interconnection in Telecommunications," *Economic Journal*, 108(448), 545–564.

——— (1999): "Competition in the Pay-TV Market," *Journal of the Japanese and International Economies*, 13(4), 257–280.

——— (2001): "Access Pricing, Bypass and Universal Service," *American Economic Review*, 91(2), 297–301.

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

Armstrong, M., C. Doyle, and J. Vickers (1996): "The Access Pricing Problem: A Synthesis," *Journal of Industrial Economics*, 44(2), 131–150.

Armstrong, M., and J. Vickers (1993): "Price Discrimination, Competition and Regulation," *Journal of Industrial Economics*, 41(4), 335–360.

——— (1998): "The Access Pricing Problem With Deregulation: A Note," *Journal of Industrial Economics*, 46(1), 115–121.

——— (2001): "Competitive Price Discrimination," *Rand Journal of Economics*, 32(4), 579–605.

Baumol, W. (1983): "Some Subtle Issues in Railroad Regulation," *International Journal of Transport Economics*, 10(1-2), 341–355.

——— (1999): "Having Your Cake: How to Preserve Universal-Service Cross Subsidies While Facilitating Competitive Entry," *Yale Journal on Regulation*, 16(1), 1–18.

Baumol, W., J. Ordover, and R. Willig (1997): "Parity Pricing and its Critics: A Necessary Condition for Efficiency in the Provision of Bottleneck Services to Competitors," *Yale Journal on Regulation*, 14(1), 145–164.

Baumol, W., and G. Sidak (1994a): "The Pricing of Inputs Sold to Competitors," *Yale Journal on Regulation*, 11(1), 171–202.

——— (1994b): *Toward Competition in Local Telephony*. MIT Press, Cambridge, MA.

Biglaiser, G., and M. Riordan (2000): "Dynamics of Price Regulation," *Rand Journal of Economics*, 31(4), 744–767.

Carter, M., and J. Wright (1994): "Symbiotic Production: the Case of Telecommunication Pricing," *Review of Industrial Organization*, 9, 365–378.

——— (1999): "Interconnection in Network Industries," *Review of Industrial Organization*, 14(1), 1–25.

——— (2003): "Asymmetric Network Interconnection," *Review of Industrial Organization*, 22(1), 27–46.

Cave, M., and M. Donnelly (1996): "The Pricing of International Telecommunications Services by Monopoly Operators," *Information Economics and Policy*, 8, 107–123.

Crew, M., and P. Kleindorfer (1998): "Efficient Entry, Monopoly, and the Universal Service Obligation in Postal Service," *Journal of Regulatory Economics*, 14(2), 103–126.

de Bijl, P., and M. Peitz (2001): "Dynamic Regulation and Competition in Telecommunications Markets - A Framework for Policy Analysis," mimeo, Netherlands Bureau for Economic Policy Analysis.

——— (2002): *Regulation and Entry into Telecommunications Markets*. Cambridge University Press, Cambridge, UK.

De Fraja, G. (1999): "Regulation and Access Pricing with Asymmetric Information," *European Economic Review*, 43(1), 109–134.

Dessein, W. (2003): "Network Competition in Nonlinear Pricing," *Rand Journal of Economics*, 34(4), 593–611.

Diamond, P., and J. Mirrlees (1971): "Optimal Taxation and Public Production: I - Production Efficiency," *American Economic Review*, 61(1), 8–27.

Doane, M., D. Sibley, and M. Williams (1999): "Having Your Cake: How to Preserve Universal Service Cross Subsidies While Facilitating Entry (Response)," *Yale Journal on Regulation*, 16(2), 311–326.

Domon, K., and K. Kazuharu (1999): "A Voluntary Subsidy Scheme for the Accounting Rate System in International Telecommunications Industries," *Journal of Regulatory Economics*, 16, 151–165.

Economides, N. (1998): "The Incentive for Non-Price Discrimination by an Input Monopolist," *International Journal of Industrial Organization*, 16(3), 271–284.

Economides, N., and L. White (1995): "Access and Interconnection Pricing: How Efficient is the Efficient Component Pricing Rule?," *The Antitrust Bulletin*, 40(3), 557–579.

Gans, J. (2001): "Regulating Private Infrastructure Investment: Optimal Pricing for Access to Essential Facilities," *Journal of Regulatory Economics*, forthcoming.

Gans, J., and S. King (2000): "Mobile Network Competition, Customer Ignorance and Fixed-to-Mobile Call Prices," *Information Economics and Policy*, 12(4), 301–328.

——— (2001): "Using 'Bill and Keep' Interconnect Agreements to Soften Network Competition," *Economics Letters*, 71(3), 413–420.

Gans, J., and P. Williams (1999): "Access Regulation and the Timing of Infrastructure Investment," *Economic Record*, 79, 127–138.

Hahn, J.-H. (2004): "Network Competition and Interconnection with Heterogeneous Subscribers," *International Journal of Industrial Organization*, 22(1), 611–631.

Hakim, S., and D. Lu (1993): "Monopolistic Settlement Agreements in International Telecommunications," *Information Economics and Policy*, 5, 147–157.

Hausman, J. (2000): "Mobile Telephony," in *Handbook of Telecommunications Economics*, ed. by M. Cave, S. Majumdar, and I. Vogelsang. North-Holland, Amsterdam, Forthcoming.

Hausman, J., and G. Sidak (1999): "A Consumer-Welfare Approach to the Mandatory Unbundling of Telecommunications Networks," *Yale Law Journal*, 109(3), 420–505.

Hermalin, B., and M. Katz (2001): "Network Interconnection with Two-Sided User Benefits," mimeo, University of California, Berkeley.

Jeon, D.-S., J.-J. Laffont, and J. Tirole (2004): "On the Receiver Pays Principle," *Rand Journal of Economics*, 35(1), 85–110.

Jorde, T., G. Sidak, and D. Teece (2000): "Innovation, Investment and Unbundling," *Yale Journal on Regulation*, 17, 1–37.

Kim, J.-Y., and Y. Lim (2001): "An Economic Analysis of the Receiver Pays Principle," *Information Economics and Policy*, 13, 231–260.

Klemperer, P. (1995): "Competition when Consumers have Switching Costs," *Review of Economic Studies*, 62(4), 515–539.

Laffont, J.-J., S. Marcus, P. Rey, and J. Tirole (2001): "Internet Interconnection and the Off-Net-Cost Pricing Principle," mimeo, Toulouse.

LAFFONT, J.-J., P. REY, AND J. TIROLE (1998a): "Network Competition: I. Overview and Nondiscriminatory Pricing," *Rand Journal of Economics*, 29(1), 1–37.

——— (1998b): "Network Competition: II. Price Discrimination," *Rand Journal of Economics*, 29(1), 38–56.

LAFFONT, J.-J., AND J. TIROLE (1994): "Access Pricing and Competition," *European Economic Review*, 38(9), 1673–1710.

——— (1996): "Creating Competition Through Interconnection: Theory and Practice," *Journal of Regulatory Economics*, 10(3), 227–256.

——— (2000): *Competition in Telecommunications*. MIT Press, Cambridge, MA.

LAPUERTA, C., AND W. TYE (1999): "Promoting Effective Competition Through Interconnection Policy," *Telecommunications Policy*, 23(2), 129–145.

LEE, S.-H., AND J. HAMILTON (1999): "Using Market Structure to Regulate a Vertically Integrated Monopolist," *Journal of Regulatory Economics*, 15(3), 223–248.

LEWIS, T., AND D. SAPPINGTON (1999): "Access Pricing With Unregulated Downstream Competition," *Information Economics and Policy*, 11(1), 73–100.

LIPSEY, R., AND K. LANCASTER (1956): "The General Theory of the Second Best," *Review of Economic Studies*, 24, 11–32.

LITTLE, I., AND J. WRIGHT (2000): "Peering and Settlement in the Internet: An Economic Analysis," *Journal of Regulatory Economics*, 18(2), 151–173.

MANDY, D. (2000): "Killing the Goose that Laid the Golden Egg: Only the Data Know Whether Sabotage Pays," *Journal of Regulatory Economics*, 17(2), 157–172.

MAYER, W. (1981): "Theoretical Considerations of Negotiated Tariff Agreements," *Oxford Economic Papers*, 33, 135–153.

POLETTI, S., AND J. WRIGHT (2003): "Network Interconnection with Participation Constraints," mimeo, University of Auckland.

REY, P., AND J. TIROLE (2006): "A Primer on Foreclosure," in *Handbook of Industrial Organization: Volume III*, ed. by M. Armstrong, and R. Porter. North-Holland, Amsterdam, forthcoming.

RIORDAN, M. (2002): "Universal Residential Telephone Service," in *Handbook of Telecommunications Economics: Volume I*, ed. by M. Cave, S. Majumdar, and I. Vogelsang. North-Holland, Amsterdam.

Sibley, D., and D. Weisman (1998): "Raising Rivals' Costs: The Entry of an Upstream Monopolist Into Downstream Markets," *Information Economics and Policy*, 10(4), 451–470.

Sidak, G., and D. Spulber (1996): "Deregulatory Takings and Breach of the Regulatory Contract," *New York University Law Review*, 4, 851–999.

——— (1997a): *Deregulatory Takings and the Regulatory Contract*. Cambridge University Press, Cambridge.

——— (1997b): "Givings, Takings and the Fallacy of Forward-Looking Costs," *New York University Law Review*, 5, 1068–1164.

Tirole, J. (1988): *The Theory of Industrial Organization*. MIT Press, Cambridge, MA.

Tye, W., and C. Lapuerta (1996): "The Economics of Pricing Network Interconnection: Theory and Application to the Market for Telecommunications in New Zealand," *Yale Journal on Regulation*, 13(2), 419–500.

Vickers, J. (1995): "Competition and Regulation in Vertically Related Markets," *Review of Economic Studies*, 62(1), 1–17.

——— (1996): "Market Power and Inefficiency: A Contracts Perspective," *Oxford Review of Economic Policy*, 12(4), 11–26.

——— (1997): "Regulation, Competition, and the Structure of Prices," *Oxford Review of Economic Policy*, 13(1), 15–26.

Weisman, D. (1995): "Regulation and the Vertically Integrated Firm: The Case of RBOC Entry Into InterLATA Long Distance," *Journal of Regulatory Economics*, 8(3), 249–266.

Willig, R. (1979): "The Theory of Network Access Pricing," in *Issues in Public Utility Regulation*, ed. by H. Trebing. Michigan State University Press, East Lansing, MI.

Wright, J. (1999a): "Competition and Termination in Cellular Networks," Mimeo, University of Auckland.

——— (1999b): "International Telecommunications, Settlement Rates, and the FCC," *Journal of Regulatory Economics*, 15, 267–291.

Yun, K., H. Choi, and B. Ahn (1997): "The Accounting Revenue Division in International Telecommunications: Conflicts and Inefficiencies," *Information Economics and Policy*, 9, 71–92.