Network Interconnection with Asymmetric Networks and Heterogeneous Calling Patterns*

Mark Armstrong
Department of Economics
University College London, UK

October 2001: this version September 2003

Abstract

This paper analysis access pricing in a telecommunications market where there is a regulated firm facing a large number of price-taking rivals. Subscribers are heterogeneous in their demand for calls. Both cases of one-way and two-way network access are analyzed. In the two-way interconnection case, a kind of duality emerges: an increase in the difference between termination charges affects the average intensity of competition, while an increase in the average termination charge affects the relative intensity of competition for the high and low volume subscribers. If the incumbent is regulated so that it just breaks even, then a reciprocal termination charge is optimal. This reciprocal charge is above the incumbent's cost of access whenever its retail tariff involves subsidizing low volume users at the expense of high volume users.

1 Introduction and Summary

This paper is about how best to regulate the terms of access to networks when (i) networks are asymmetric in terms of demand and/or costs and (ii) subscribers are heterogeneous in their demand for calls. The setting is one involving a regulated incumbent firm which faces a large number (a "competitive fringe") of price-taking entrants. The analysis is divided into two sections, the first of which examines the case of so-called "one way" access pricing, while the second looks at "two way" interconnection.

A summary of the results is as follows. (The derivations of these results are found in the main body of the paper.)

*Funding assistance from the Chilean government is gratefully acknowledged. I am also grateful to Wouter Dessein and Martin Peitz for comments on an earlier draft.

One way access pricing with heterogeneous subscribers: The situation is one where an incumbent firm offers service to two groups of subscribers (high-volume and low-volume subscribers) at regulated terms. It might be that the incumbent's tariff fully reflects the underlying costs; more likely in practice, a degree of cross-subsidy from high users to low users will be present in its tariff. Entrants require access to the incumbent's network to offer their own service. There is an increasing relationship between the incumbent's access charge and the retail prices offered by entrants.

In the case where the incumbent's retail tariff fully reflects its costs, the optimal access charge regime is simple: entrants should have access to the incumbent's network at cost. Because there are no distortions in the incumbent's retail tariff, there is no reason to distort the entrants' use of the incumbent's network.

When the incumbent's tariffs involves excess profits being generated from high users and losses from low users, then a cost-based access regime will lead to inefficient "cream-skimming" entry into the high-user segment, as well as too little entry to the low-user segment. Ideally, regulation should implement above-cost access to the network when an entrant serves a high-volume subscriber, and below-cost access when an entrant serves a low-volume subscriber. However, this form of "price discrimination" might not be feasible, in which case a compromise is inevitable: the optimal *uniform* access charge will lie between the ideal charge for high users and the ideal charge for low users, and an inefficiency pattern of entry will result. Essentially, there is a problem of too few regulatory instruments (here, a single, uniform access charge) to achieve the regulatory objectives (here, obtaining the desirable degree of entry into each of the two market segments).

Two way network interconnection: In essence, unlike the one-way case where calls were just "made", here calls are both "made" and "received". In particular, entrants as well as the incumbent receive revenue from providing network access, and the revenue that entrants receive for terminating calls *to* their subscribers affects their competitive strategy.

In this case, the relationship between termination charges and the entrants' retail prices is as follows:

- Increasing the incumbent's termination charge causes entrants to raise their prices to all subscribers. This is analogous to raising the (one way) access charge as just discussed.

- Increasing the entrants' termination charge will cause entrants to lower their prices to all subscribers (provided entrants do not have a large aggregate market share). This is because when entrants obtain higher revenue from terminating calls to their own subscribers, this makes it more profitable to attract further subscribers, and so competition is intensified.

- Increasing both termination charges equally (in particular, increasing a *reciprocal* termination charge) will cause entrants to lower their price to low-volume users and to raise their price to high-volume users. The reason for this is that an increase in *both* termination charges makes it relatively more profitable for an entrant to attract a subscriber with a net inflow of calls (and relatively less profitable to attract a subscriber who makes more calls than he receives). In the specific model considered below, it is the low-volume subscribers who have a net inflow of calls, and so entrants compete harder for these subscribers. (In the special case where subscribers are homogeneous, increasing both termination charges equally has no effect at all on the entrants' strategy — see expression (22) below.)

In rough terms, then, a kind of "duality" emerges, and an increase in the *difference* between termination charges on the two networks affects the *average* intensity of competition for the two groups of subscribers, while an increase in the *average* termination charge affects the *relative* intensity of competition for the two groups of subscribers. This insight drives the main results about the desirable pattern of termination charges:

1. If the incumbent is regulated so that it just breaks even overall, then a reciprocal termination charge is optimal. (See expression (20) below.) If the incumbent made a positive profit on average, then, all else equal, it would be optimal to *inhibit* competition to some degree in order to preserve socially valuable profit, and this is done by allowing the incumbent to charge more for call termination than the entrants.

2. If the incumbent is required to cross-subsidize low-volume users from profits earned from high-volume users (but just breaks even overall), then it is desirable to set a *high* (reciprocal) termination charge, in the sense that the charge should be above the incumbent's cost of termination. (See expression (24) below.) The reason is that this choice of termination charge makes entrants compete relatively aggressively for low-volume subscribers (i.e., subscribers with a net inflow of calls), which is desirable given that the incumbent is making a loss on these subscribers.

2 Related Literature

There is a small, but expanding, list of papers that examine issues related to this paper. One strand of work examines the socially optimal choice of access charges when networks are asymmetric in terms of demand (but not cost) and are unregulated at the retail level. Carter and Wright (2003) analyze the networks' preferences over a *reciprocal* access charge. They show that the network which is preferred by more subscribers (the "large network") will wish to set a reciprocal access charge equal to the cost of termination. Moreover, they show that if the "small" network is sufficiently small, then it too will prefer cost-based access pricing. (However, in their model if the networks are very asymmetric then cost-based access pricing is not socially optimal, even among the class of reciprocal access charging regimes.) Peitz (2003) examines the effect of making access charges different for the two networks, i.e., of non-reciprocal access charges. (The model is otherwise essentially the same as Carter and Wright (2003) and, in particular, there are no cost differences between networks.) To be precise, Peitz discusses the effect of increasing the small network's access charge above cost while keeping the larger network's access charge equal to cost. He finds that, compared to the case where both networks have cost-based access, such a policy causes the entrant's profit to rise, subscribers to be better off, and for total welfare (as conventionally measured) to decline.[1]

Pages 374 to 377 of Armstrong (2002) focus more on the effects of cost asymmetries.[2] The analysis is greatly simplified by the assumption that subscribers' demand for calls is completely inelastic, an assumption that will be maintained in the current paper. In this framework I show that the retail tariffs offered by *both* networks decrease if the small network can charge more for access than the large network. (With inelastic demand, the average level of access charges has no effect on equilibrium retail tariffs, and only the difference between the two networks' access charges has any effect.) The profit of the small network increases (and the profit of the large network decreases) with the difference between the two networks' access charge. (Clearly, these comparative statics results are consistent with the results derived in Peitz (2003).)

The previous paragraphs discussed the framework in which there were two networks which were free to choose their retail tariffs, given the access charges. Other sections of Armstrong (2002) examine cases where one or more firms are regulated at the retail level. Sections 4.2.1 and 4.2.2 of Armstrong (2002) suggest a framework for discussing competition between asymmetric networks and analyze the socially optimal distribution of consumers across these two networks. A conclusion of this analysis was that (regulated) retail tariffs should be adjusted away from natural measures of underlying cost in order to encourage consumers to choose the network with the lower *termination* cost, all else equal. (When tariffs are based on the cost of making, but not receiving, calls, then there is an negative externality when a subscriber joins a network with a high termination cost, and adjustments should be made so that subscribers internalize this externality.)

Pages 377 to 379 in Armstrong (2002) introduce a highly stylized model of a regulated incumbent which faces a number of (unregulated) entrants. These entrants are all identical, but offer a service that is differentiated from that of the incumbent. There, because of the assumption of inelastic demand, only the difference between the two access charges matters (and the average level has no effect on equilibrium outcomes). The entrants' equilibrium retail tariff is a decreasing function of the difference between their access charge and the incumbent's. The main result is that, if the incumbent is regulated so that it makes positive

[1] However, chapters 3 and 4 of de Bijl and Peitz (2002) argue that this overall loss in welfare is likely to be "small".

[2] Section 6.3 of de Bijl and Peitz (2002) contains similar analysis.

profits on its service, it is socially desirable for the incumbent to charge more for access than the entrants. (We will see in section 4 that this result carries over to the case of heterogeneous subscribers.) The reason is that this policy makes entrants compete less vigorously for subscribers, and thus it protects the socially valuable profit of the regulated incumbent.

The current paper extends the framework in pages 377 to 379 of Armstrong (2002) to the case where subscribers have different demand for calls (some are high-volume users and some are low-volume users). As such, previous work on competition between networks with nonlinear pricing is relevant. The two key contributions in this area are Dessein (2002) and Hahn (2002).[3] These two papers assume symmetric competition and reciprocal access charges, and investigate how the level of the access charge affects the tariffs offered by firms in equilibrium. One insight — see Proposition 2 in Hahn (2002) for instance — is that, when the access charge is above cost, networks will compete relatively hard for subscribers who generate a net *inflow* of calls. In natural cases, it is the low-volume users who make fewer calls than they receive, and so a high reciprocal access charge will induce networks to offer low retail tariffs targeted as such subscribers. This insight plays a crucial role in the analysis that follows.

3 One way access pricing

The following analysis is closely parallel to that in section 4, and one purpose is to provide a gentle introduction to the more complicated analysis of two-way interconnection in markets with high- and low-volume subscribers.[4] The situation considered here is that there is an incumbent firm, network A, which has infrastructure in place that is needed to provide service to consumers. Entrants can provide a (differentiated) service to consumers, provided that entrants obtain suitable access to the incumbent's network. Entrants, denoted "network B" for simplicity, are many in number, provide identical services (which is differentiated from the incumbent's) and act as price takers. This modeling of entrants as a "competitive fringe" is done mainly for analytical tractability, although in many cases it is not such a bad approximation in practice.[5]

A common situation is for the incumbent firm to be required to offer a retail tariff that favors low-volume users at the expense of high-volume users. A major question for policy is how to design access charges in order to prevent entrants "cream-skimming" the profitable high-volume users. To conceptualize the issues involved here, suppose that there are two

[3] See Poletti and Wright (2003) for an extension of this work to the case where some subscribers' participation constraints are binding.

[4] For a closely related analysis, see section 3.2.3.3 of Laffont and Tirole (2000).

[5] 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 the incumbent's 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.

groups of subscribers, and a *high* (respectively *low*) demand type of subscriber is denoted by H (L). Suppose that subscribers have inelastic demand for calls, and the type d subscriber necessarily makes X^d calls, where $d = L, H$. The fraction of subscribers with high demand is α. We assume that all subscribers choose to join one or other network, so that the total number of subscribers is constant in the analysis, and this total number is normalized to 1.

Recall that the two networks, respectively A and B, are the incumbent and the competitive fringe. A network will offer a pair of contracts, one for each of the two kinds of subscriber, and suppose that the charge for the X^d calls made by a type-d subscriber on network $i = A, B$ is P_i^d. For simplicity suppose that firms can directly observe the type of a given subscriber, so that a tariff P_i^d can be explicitly targeted at the demand type. (Using the standard terminology, suppose the firms can practice third-degree price discrimination rather than having to rely on second-degree price discrimination.) Therefore, the number of type d subscribers attracted to network i depends only on the pair of prices $\{P_A^d, P_B^d\}$ and not on the prices offered to the other type of subscriber. Services supplied by the two networks are differentiated, and if $\{P_A^d, P_B^d\}$ are the two prices offered to the type d subscribers, the fraction of these subscribers who obtain service from network i is n_i^d (which is some function of these prices).

The incumbent's unit cost of providing its retail service is C_1, its unit cost of providing access to the entrants is C_2, while the entrant's unit cost of providing its retail service (given that it has one unit of access from the incumbent per unit of its output) is c. Each network $i = A, B$ has fixed cost k_i per subscriber when providing service. Write $C_A^d = k_A + C_1 X^d$ for the total costs incurred when the incumbent supplies the service to a type d subscriber, and let $C_B^d = k_B + (C_2 + c)X^d$ be the total physical costs incurred when an entrant supplies service to a type d subscriber.

Suppose network i attracts a fraction n_i^d of the type d subscribers. Since access charges are just transfers between the firms, total industry profits $\Pi_A + \Pi_B$ are

$$\Pi = \alpha n_A^H\{P_A^H - C_A^H\} + (1-\alpha)n_A^L\{P_A^L - C_A^L\} + \alpha n_B^H\{P_B^H - C_B^H\} + (1-\alpha)n_B^L\{P_B^L - C_B^L\} .$$

Socially desirable retail tariffs: Suppose the incumbent firm A has a regulated retail tariff $\{P_A^L, P_A^H\}$. Using familiar arguments, the optimal tariff offered by the competitive fringe B, given the incumbent's tariff, is given by the "equal markup" rule[6]

$$P_B^d - C_B^d = P_A^d - C_A^d \quad \text{for } d = L, H . \tag{1}$$

In particular, if A is optimally regulated, in the sense that $P_A^d = C_A^d$ for both types of subscriber, then this expression (1) implies that the competitive fringe B should also have its retail prices set equal to its physical costs of supply.

[6]For instance, see expression (90) in Armstrong (2002). This equal markup rule depends on the assumption that the total number of active subscribers is constant over the relevant range of prices. In this case, total welfare depends only upon the *difference* between the two networks' prices, and the social optimum involves the price difference being set equal to the difference in underlying costs.

The relationship between the access charge and the entrants' retail tariff: Suppose the incumbent's access charge is a, and this is required to be the same regardless of the type of the subscriber (i.e., whether the access is used to supply service to a high- or a low-volume user). Competition within the fringe will mean that each fringe firm will choose a retail charge that just covers its total costs. In other words, the equilibrium tariff for a fringe firm offering service to a type d subscriber is

$$P_B^d = k_B + (c+a)X^d = C_B^d + (a - C_2)X^d \ . \tag{2}$$

In particular, when access is priced at cost ($a = C_2$) we have $P_B^d = C_B^d$ in equilibrium. More generally, the entrants' tariff is increasing in the access charge, and an increase in a causes both P_B^L and P_B^H to increase and the incumbent to gain market share for both subscriber groups. This result contrasts with the two-way case analyzed in the next section, where a rise in (all) access charges acts to make entrants compete more aggressively for high users and to compete less aggressively for low users.

Using access charges to implement the desirable outcome: We next turn to the question of which access charging regime (if any) can implement the socially optimal outcome described in (1). Suppose for now that it *is* possible to set a different access charge for service to the two kinds of subscribers, and say a^d is the access charge if the entrant serves a type d subscriber. Given the incumbent's regulated tariffs $\{P_A^L, P_A^H\}$, expressions (1) and (2) imply that the ideal pair of access charges is given by

$$a^d = C_2 + \frac{P_A^d - C_A^d}{X^d} \ , \ d = L, H \ . \tag{3}$$

This formula is an instance of the 'ECPR' which, in broad terms, states that the access charge should be determined by the direct cost of providing access (C_2 in this case) plus the incumbent's lost profit per call when it loses a subscriber to a rival (which is $\frac{P_A^d - C_A^d}{X^d}$ in this case).[7]

The only circumstance when these two access charges are equal is if the incumbent makes the same average profit per call from each kind of subscriber, i.e., if

$$\frac{P_A^L - C_A^L}{X^L} = \frac{P_A^H - C_A^H}{X^H} \ .$$

In particular, if the incumbent's regulated retail prices just cover its associated costs for each type of subscriber, then (3) implies that both access charges should be equal to cost: $a^L = a^H = C_2$. This case is relevant if the incumbent's tariff has been fully "rebalanced" prior to allowing entry.

[7] For more detail on the ECPR, see section 2 of Armstrong (2002).

Incumbent's tariff favors low users at expense of high users: However, instead of the full rebalancing just discussed, a common situation is for the incumbent to cross-subsidize low users (e.g., residential users) out of profits generated from high users (e.g., business users). Therefore, suppose that the incumbent's regulated tariff takes the form

$$P_A^L < C_A^L \; ; \; P_A^H > C_A^H \; . \tag{4}$$

Then expression (3) states that, if differential access pricing is feasible, then the access charge for service for the high volume users should be above cost, while the access charge for service to low users should be below cost. By such means the regulator can discourage excessive and inefficient "cream-skimming" entry into the market for high volume users, and can encourage what would otherwise be inefficiently small-scale entry into the loss-making market for low users.

However, it may not be feasible to charge differentially for access to the two groups. Consider the case where both access charges are set equal to the incumbent's cost, C_2. Then, from (2), the fringe will offer the cost-based tariff $P_B^d = C_B^d$. Therefore, from (1) and (4) we see that a cost-based access charge causes the entrants to compete too aggressively for high users (P_B^H is too low) and not to compete hard enough for low users (P_B^L is too high). When the regulator is required to choose a uniform access charge for the two kinds of subscribers, the optimal uniform charge will lie somewhere between the two charges in (3). Whether the optimal uniform access charge is above or below the cost C_2 is ambiguous, and will on the relative sizes of the two groups of subscribers, the incumbent's profit margins $P_A^d - C_A^d$, and also on how responsive the two groups are to price differences between the two networks.

Regardless of the precise details, though, the result will be a compromise: there will be too much entry into the high-volume market and too little into the low-volume market. The reason is the usual one: there are too few instruments to achieve the required number of objectives. In this instance, there is a single instrument (the uniform access charge) and two objectives (to achieve the desirable amount of entry into each of the two market segments of high users and low users).

4 Two way network interconnection

4.1 Basic framework

Here, we modify the simple framework in section 3 to allow subscribers to make and receive calls. Suppose, as before, subscribers differ in the volume of calls they make. In the two-way framework they might also differ in the number of calls they receive. A *high* (respectively *low*) demand type of subscriber is denoted by H (L). As before, subscribers have inelastic demand for calls, and the type d subscriber necessarily makes X^d calls. Subscribers are sorted according to their volume of outbound calls, so that $X^L \leq X^H$. Suppose that a fraction β^d of the calls made by type d subscribers are made to type H subscribers, and the fraction $1-\beta^d$ of their calls are made to low demand consumers.[8] (If $\beta^d \equiv \alpha$ then we would have a model where subscribers were equally likely to call every other subscriber, regardless of the demand characteristics of the called, or calling, subscriber.)

The number of calls received by a type d subscriber, denoted Y^d, therefore, is

$$Y^L = \frac{\alpha}{1-\alpha}(1-\beta^H)X^H + (1-\beta^L)X^L \; ; \; Y^H = \beta^H X^H + \frac{1-\alpha}{\alpha}\beta^L X^L \; .$$

(Perhaps the main advantage of making the strong assumption of inelastic demand for calls is that the volume of incoming calls for a given subscriber does not depend on the contracts offered by the firms, nor on the market shares.) Clearly, the total number of calls made equals the total number of calls received:

$$\alpha X^H + (1-\alpha)X^L = \alpha Y^H + (1-\alpha)Y^L \; . \tag{5}$$

The type H subscribers make more calls than they receive ($X^H > Y^H$) provided that

$$(1-\alpha)\beta^L X^L < \alpha(1-\beta^H)X^H \; ,$$

in which case the type L subscribers make fewer calls than they receive. This is probably the more natural of the two cases to consider. (In particular, it is satisfied if $\beta^d \approx \alpha$.)

As before, a network will offer a pair of tariffs, one for each of the two kinds of subscriber, and suppose that the combined charge for all X^d calls made by a type d subscriber on network $i = A, B$ is P_i^d.[9] Again, suppose for simplicity that firms can observe the type of a given subscriber, so that a tariff P_i^d can be explicitly targeted at the demand type. Therefore, the number of type d subscribers attracted to network i depends only on the pair of prices $\{P_A^d, P_B^d\}$ and not on the prices offered to the other type of subscriber. The termination charge on network i is a_i, and this is required to be the same regardless of the types of the caller or recipient. However, we do not impose a restriction that $a_A = a_B$ (i.e., that termination charges are "reciprocal"), although this will turn out to be socially desirable in some special cases.

Suppose network i has a fraction n_i^d of the type d subscribers. Write Q_{ij} for the total number of calls made from network i to network j. Then the total profit of network i is

$$\alpha n_i^H\{P_i^H - k_i\} + (1-\alpha)n_i^L\{P_i^L - k_i\} - Q_{ii}(c_i^O + c_i^T) - Q_{ij}(c_i^O + a_j) + (a_i - c_i^T)Q_{ji}$$

[8] The proposed model assumes that the number of calls received by a subscriber is perfectly correlated with the number of calls made. A more attractive, albeit more complicated, model might have *four* types of subscriber: a subscriber could have high or low demand for outbound calls, and a subscriber could be a person whom others wish to call a little or a lot.

[9] We assume that firms commit to make the combined charge independent of realized market shares. If, for instance, firm A set one price for on-net calls (calls within its network) and a different price for calls to network B, then subscribers will care about the realized market shares when they decide which network to join (because that will affect the total charge for their calls). These network effects will make the analysis more complicated, and are ignored in this paper.

$$= \alpha n_i^H \{P_i^H - k_i\} + (1-\alpha) n_i^L \{P_i^L - k_i\} - (Q_{ii} + Q_{ij}) c_i^O - (Q_{ii} + Q_{ji}) c_i^T + a_i Q_{ji} - a_j Q_{ij} \ .$$

(Here, c_i^O is network i's cost of originating a call, c_i^T is its cost of terminating a call, and k_i is its fixed cost for supplying service to a subscriber.) Notice that $Q_{ii} + Q_{ij}$ is the total number of calls originating on i's network, and so

$$Q_{ii} + Q_{ij} = \alpha n_i^H X^H + (1-\alpha) n_i^L X^L \ . \tag{6}$$

Similarly, $Q_{ii} + Q_{ji}$ is the total number of calls terminated on network i, and so

$$Q_{ii} + Q_{ji} = \alpha n_i^H Y^H + (1-\alpha) n_i^L Y^L \ . \tag{7}$$

Write

$$C_i^d = c_i^O X^d + c_i^T Y^d + k_i$$

for the total physical costs incurred by network i when providing service to a type d subscriber. (These cost parameters play a central role in the following, and it is important to bear in mind that they include the cost associated with delivering calls *to* the subscriber, and they exclude the costs incurred by network j when it delivers network i's calls.) With this notation, network i's profits can be written simply as

$$\Pi_i = \alpha n_i^H \{P_i^H - C_i^H\} + (1-\alpha) n_i^L \{P_i^L - C_i^L\} + a_i Q_{ji} - a_j Q_{ij} \ .$$

Therefore, since termination payments cancel out, total industry profits $\Pi_A + \Pi_B$ are

$$\Pi = \alpha n_A^H \{P_A^H - C_A^H\} + (1-\alpha) n_A^L \{P_A^L - C_A^L\} + \alpha n_B^H \{P_B^H - C_B^H\} + (1-\alpha) n_B^L \{P_B^L - C_B^L\} \ .$$

Finally, for use later it is convenient to introduce the following notation: let

$$s_i^d = \beta^d n_i^H + (1-\beta^d) n_i^L \ , \tag{8}$$

so that s_i^d is the fraction of calls made by all of the type-d subscribers that are received on network i.

4.2 Socially desirable retail tariffs

Suppose the incumbent, network A, offers the regulated retail tariff $\{P_A^L, P_A^H\}$. For the same reasons as in section 3, the optimal tariffs offered by network B satisfy the equal markup rule

$$P_B^d - C_B^d = P_A^d - C_A^d \quad \text{for } d = L, H \ . \tag{9}$$

In the special case where the costs for each network are identical, so that $C_A^d = C_B^d$, then the two networks should charge the same prices, so that $P_A^d = P_B^d$.

Expression (9), and the particular notion of "cost" that is relevant, is central to a good understanding of competition between asymmetric networks. A naive view might be to base prices on other measures of costs. For instance, suppose that network A's retail tariff is set equal to the total *outbound* costs of making calls from its network (with no contribution to inbound costs). Specifically, suppose that

$$P_A^d = k_A + X^d(c_A^O + \bar{c}^d)$$

where $\bar{c}^d = s_A^d c_A^T + s_B^d c_B^T$ is the average termination cost on the two networks, given the market shares for receiving calls as defined in (8). Then (9) implies that the optimal retail price for network B is

$$P_B^d = \underbrace{k_B + X^d(c_B^O + \bar{c}^d)}_{\text{outbound call costs}} + \underbrace{Y^d(c_B^T - c_A^T)}_{\text{adjustment factor}} \ . \tag{10}$$

Thus, as discussed in section 2, given that network A has a tariff based on its outbound call costs, it is generally not socially optimal for firm B to do likewise, and B's tariff should be adjusted up (or down) if its termination costs are higher (or lower) than A's. Thus, it is socially desirable to distort retail tariffs away from the underlying costs of *outbound* calls in order to drive subscribers away from the network with the higher termination cost. If tariffs are based purely on outbound calling costs, then subscribers ignore the externality that their choice of network imposes on their callers.

Clearly, though, if networks do not differ significantly in their termination costs ($c_A^T \approx c_B^T$), then the adjustment factor in (10) vanishes, and so basing both networks' retail tariffs on their respective outbound calling costs is one way to implement the socially desirable outcome.

4.3 Relationship between termination charges and retail tariffs

As in section 3, suppose that network B is composed of a large number of small but identical firms. Therefore, competition within the fringe will mean that each entrant will choose a retail charge that just covers its total costs. Suppose that a fringe firm manages to attract a type d subscriber. If it charges the retail price P_B^d, its total profits from this subscriber are

$$\Pi_B^d = P_B^d - k_B - X^d[c_B^O + s_A^d a_A + s_B^d a_B] + (a_B - c_B^T)Y^d \ . \tag{11}$$

(Here we assumed that the firm was "small" in the sense that a negligible fraction of calls made by its subscribers were received by its subscribers.) Therefore, since the fringe firms must just break even when serving either type of subscriber, we see that the equilibrium fringe tariffs are given by

$$P_B^d = C_B^d + X^d[s_A^d a_A + s_B^d a_B] - Y^d a_B \quad \text{for } d = L, H \ . \tag{12}$$

We can re-write these equilibrium prices in terms of a_B and difference in termination charges $\Delta_a = a_A - a_B$:

$$P_B^d = C_B^d + (X^d - Y^d)a_B + s_A^d X^d \Delta_a \ . \tag{13}$$

(It is important to recognize that s_A^d in the above expression depends positively on both fringe prices, P_B^L and P_B^H.)

One natural case for consideration is when the termination charges are set equal to termination costs, so that $a_i = c_i^T$. In this case (12) implies that

$$P_B^d = k_B + X^d(c_B^O + \bar{c}^d) \; , \tag{14}$$

where $\bar{c}^d$ is again the average cost of call termination given the market shares of the two networks. Therefore, if termination charges are equal to costs then the equilibrium tariff of the entrants is equal to the actual physical cost of making outbound calls.

Alternatively, another natural case is the so-called "bill-and-keep" regime, where $a_A = a_B = 0$. In this case, (13) implies that network B has cost-based pricing: $P_B^d = C_B^d$. More generally, suppose there is a reciprocal charging regime, so that $a_A = a_B = a$, say. Then (13) simplifies to

$$P_B^L = C_B^L + \left[X^L - Y^L\right] a \; ; \; P_B^H = C_B^H + \left[X^H - Y^H\right] a \; . \tag{15}$$

Therefore, prices are distorted from the underlying costs (as measured by C_B^d) according to whether the subscriber makes more, or fewer, calls than he receives. As we have said, the typical case is where $X^H > Y^H$ and $X^L < Y^L$, in which case (15) implies that the equilibrium fringe price for the low-volume subscribers falls with a, while the fringe price for high-volume subscribers rises with a. This is quite intuitive: a high reciprocal termination charge will discourage competition for those subscribers who make more calls than they receive, since such subscribers generate a loss on net termination payments. Put another way, a low reciprocal termination charge—such as the bill-and-keep system—will encourage entrants to compete aggressively for those subscribers with a net outflow of calls (the high-volume users in this model).[10]

Another way to gain insight into the effect of the choice of (reciprocal) termination charge is to consider the incumbent's net outflow of calls. From expressions (6) and (7), this net outflow is

$$\text{incumbent's net outflow} = \alpha n_A^H(X^H - Y^H) + (1-\alpha) n_A^L(X^L - Y^L) \; . \tag{16}$$

Since, in the usual case, expression (15) implies that increasing a causes P_B^L to fall and P_B^H to rise, it follows that increasing a causes n_A^L to fall and n_A^H to rise. Therefore, from (16) it follows that raising a causes the incumbent's outflow of calls to increase. Of course, this is just another way of saying that high termination charges induce the entrants to compete aggressively for low-volume users (who have a net inflow of calls), but it serves to emphasize that with heterogeneous calling patterns there is no reason to expect balanced

[10]More generally, an equal shift upwards in both termination charges has the same effect, and will cause the entrants' low-volume price to fall and their high-volume price to rise.

call flows across networks, and the termination charge acts systematically to affect aggregate call flows.

Next, consider the effect of raising a_A, keeping a_B fixed.[11] Expression (13) indicates that the prices offered by the entrants increase (for both demand types), and so the incumbent will gain market share for both types. On the other hand, from (13) raising a_B, while keeping a_A fixed, will cause the price P_B^d to fall whenever $s_B^d X^d - Y^d < 0$. This inequality will always hold for the low-volume subscribers (since $X^L < Y^L$), and it will also hold for the high-volume subscribers whenever the entrants have a low enough market share of that group. In such cases, raising a_B on its own will cause both prices P_B^d to fall.

In sum, we have the following simple predictions about the relationship between termination charges and entrants' prices:

- Increasing the incumbent's termination charge a_A will cause entrants to raise their prices to all subscriber groups. This case is analogous to raising the (one way) access charge in section 3 above — see expression (2).

- Increasing the entrants' termination charge a_B will cause entrants to lower their prices to all consumer groups (provided entrants do not have a large aggregate market share). This is because when entrants obtain higher revenue from terminating calls to their subscribers, this makes it more profitable to attract further subscribers, and so competition is intensified.

- Increasing both termination charges equally will cause entrants to lower their price to low-volume users and to raise their price to high-volume users (under the maintained assumption that $X^H > Y^H$ and $X^L < Y^L$).

4.4 Using termination charges to implement the desirable outcome

We next turn to the question of which termination regime (if any) can implement the socially optimal outcome described in (9). Given the incumbent's regulated tariffs $\{P_A^L, P_A^H\}$, expression (13) implies that the ideal termination charges should satisfy

$$(X^d - Y^d)a_B + s_A^d X^d \Delta_a = P_A^d - C_A^d\ , \quad d = L, H\ . \tag{17}$$

This is the central formula in this paper. It states two equations in two unknowns, a_B and Δ_a. Some arithmetic manipulations yield the more explicit formulas:

$$\left(\frac{X^H - Y^H}{s_A^H X^H} - \frac{X^L - Y^L}{s_A^L X^L}\right) a_B = \frac{P_A^H - C_A^H}{s_A^H X^H} - \frac{P_A^L - C_A^L}{s_A^L X^L}\ ; \tag{18}$$

[11] The following comparative statics assume that Δ_a is 'small', so that we can ignore the effect of changes to s_A^d in expression (13) when we change a_A.

$$\Delta_a = \frac{\bar{\Pi}_A}{\bar{Y}} \ . \tag{19}$$

In expression (19), $\bar{\Pi}_A = (1-\alpha)[P_A^L - C_A^L] + \alpha[P_A^H - C_A^H]$ is one measure of the incumbent's "average profit" across the two groups of subscribers, and $\bar{Y} = (1-\alpha)s_A^L X^L + \alpha s_A^H X^H$ is the total number of calls terminated on network A. The term $(\cdot)$ on the left-hand side of (18) is positive in the typical case ($X^H > Y^H$), while the right-hand side is related to the difference in profitability for the incumbent serving the two groups of subscribers.

Expression (19) gives a straightforward criterion for deciding which network should have the higher termination charge. If the incumbent is profitable on average, in the sense of $\bar{\Pi}_A > 0$, then it should charge more for termination than the fringe. This makes sense following the discussion at the end of section 4.3: all else equal, a high value for Δ_a translates into high retail prices being offered by the entrants, and this means that the incumbent does not lose too much market share. (Losing too much market share is socially costly when the incumbent is earning positive profits.) In addition, there is the case where reciprocal termination charges are optimal:

$$\bar{\Pi}_A = 0 \Longrightarrow a_B = a_A \tag{20}$$

so that when the incumbent is regulated so that its average price is equal to its average costs, then a reciprocal termination charge is optimal (but not otherwise). In this case, (17) implies that the reciprocal termination charge is given by

$$a_A = a_B = \frac{P_A^L - C_A^L}{X^L - Y^L} \ \left(= \frac{P_A^H - C_A^H}{X^H - Y^H} \right) \ . \tag{21}$$

Some natural special cases include the following:

Homogenous subscribers:[12] Suppose that all subscribers make and receive the same number of calls, so that

$$X^L = X^H = Y^L = Y^H = X \ .$$

Write

$$C_i = k_i + X(c_i^O + c_i^T) \ , \ \ i = A, B \ .$$

Then (13) implies that the equilibrium retail charge offered by the fringe is

$$P_B = C_B + \Delta_a X n_A \ . \tag{22}$$

This has the feature that B's equilibrium retail price is an increasing function of Δ_a. The reason for this is that 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. Also worth emphasizing in this special case is that outcomes depend only on the difference in termination charges Δ_a,

[12]This discussion essentially replicates pp. 377–379 of Armstrong (2002).

and absolute levels do not matter. This insight reinforces the discussion in section 4.3: the difference Δ_a effects the impact of entry for all subscriber groups, whereas absolute levels of termination charges affect the relative strength of the entrants' competitive impact across the different subscriber groups. In the case of homogeneous subscribers, the second factor plays no role.

Then expression (19) implies that the optimal difference in the termination charges is

$$\Delta_a = \frac{P_A - C_A}{X n_A} \ . \tag{23}$$

(Clearly, the incumbent's market share n_A is determined endogenously in this formula.) In particular, a reciprocal termination charge ($\Delta_a = 0$) is optimal only when $P_A = C_A$. Perhaps counter-intuitively, then, when there are no regulated distortions at the retail level (in the sense that $P_A = C_A$), it is *not* the case that a (in general, non-reciprocal) cost-based termination regime, where $a_i \equiv c_i^T$, is optimal.

In other cases, however, non-reciprocal termination charges given by (23) are used to implement the optimal fringe price. If the incumbent is profitable in this market ($P_A > C_A$) then it should charge more for call termination than the fringe ($\Delta_a > 0$), regardless of the underlying relative costs of call termination.

Optimal regulation of incumbent: Here, return to the case of heterogeneous subscribers, and suppose that the incumbent's regulated retail prices just cover its associated costs—as measured by C_A^d—for each subscriber type, so that $P_A^L = C_A^L$ and $P_A^H = C_A^H$. (Recall, however, that this cost includes the costs of *delivering* all calls to the type k subscriber, and so departs from the usual measures of costs based on outbound calls. In effect, this case arises only if the incumbent charges its subscribers for *receiving* calls, and at a rate equal to its cost of terminating calls.[13]) In this case, expressions (18) and (19) imply that the optimal access charge regime is

$$a_A = a_B = 0 \ .$$

This is the "bill-and-keep" rule, where firms do not charge each other for terminating calls.

The impact of cross-subsidies: For simplicity, suppose that the incumbent breaks even overall, in the sense that $\bar{\Pi}_A = 0$ in expression (19), which implies that reciprocal termination is optimal, and so write $a = a_A = a_B$. Assuming that $X^H < Y^H$ as usual, expression (21) implies that

$$a > c_A^T \Leftrightarrow P_A^H > k_A + (c_A^O + c_A^T) X^H \ . \tag{24}$$

This formula (24) states that the reciprocal termination charge should be above the incumbent's termination cost whenever the incumbent's tariff for high-volume users exceeds

[13] See Jeon, Laffont, and Tirole (2003) for an analysis of the case where networks charge subscribers for receiving calls.

$k_A + (c_A^O + c_A^T)X^H$. This latter quantity is the hypothetical cost for the incumbent in supplying the high-volume user with outbound calls that all terminate on the incumbent's network. This is perhaps a fairly common situation, where low users are subsidized from the profit generated from high users.

The reason why a high reciprocal termination charge is optimal in this case of cross-subsidy is that the policy discourages the entrants from inefficiently attracting the profitable high users (since high users make more calls than they receive, and a high reciprocal termination charges acts to make these subscribers less attractive — see expression (15) above). Thus, a high reciprocal termination charge acts to prevent inefficient "cream-skimming" entry into the high-volume market, just as the ECPR rule in (3) does in the one-way context.

In the knife-edge case where the incumbent's retail tariff is precisely equal to its own costs for making (on-net) calls, so that

$$P_A^d = k_A + (c_A^O + c_A^T)X^d \ , \ d = L, H$$

then

$$a_A = a_B = c_A^T$$

and the optimal regime involves reciprocal termination charges at the level of the incumbent's cost. This case is relevant if the incumbent's tariff has been fully rebalanced to reflect its (outbound) costs.

References

Armstrong, M. (2002): "The Theory of Access Pricing and Interconnection," in *Handbook of Telecommunications Economics: Volume I*, ed. by M. Cave, S. Majumdar, and I. Vogelsang. North-Holland, Amsterdam.

Carter, M., and J. Wright (2003): "Asymmetric Network Interconnection," *Review of Industrial Organization*, 22, 27–46.

de Bijl, P., and M. Peitz (2002): *Regulation and Entry into Telecommunications Markets.* Cambridge University Press, Cambridge, UK.

Dessein, W. (2002): "Network Competition in Nonlinear Pricing," *Rand Journal of Economics*, forthcoming.

Hahn, J.-H. (2002): "Network Competition and Interconnection with Heterogeneous Subscribers," mimeo, Keele University.

Jeon, D.-S., J.-J. Laffont, and J. Tirole (2003): "On the Receiver Pays Principle," *Rand Journal of Economics*, forthcoming.

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

Peitz, M. (2003): "Asymmetric Access Price Regulation in Telecommunications Markets," *European Economic Review*, forthcoming.

Poletti, S., and J. Wright (2003): "Network Interconnection with Participation Constraints," mimeo, University of Auckland.