Competition in communication networks:

Pricing and regulation

Robin Mason, University of Southampton and CEPR

and

Tommaso M. Valletti, Imperial College and CEPR

A paper prepared for the *Oxford Review of Economic Policy*

Issue on European Network Infrastructures

1. Introduction

By any measure, the communication sector is important. The combined revenues of the U.S. telephone and postal industries are larger than military spending for that country, and almost three times higher than the revenues of the airline industry. (Table 1, taken from Odlyzko (2000), gives further comparisons.) The sector has been expanding for many centuries. The postal service accounted for 0.02% of the U.S. gross domestic product (GDP) in 1790, now it accounts for around 0.7%; the number of pieces of mail per person per year has risen from 0.2 to 733 over the same period. The telephone industry has grown from 0.12% of U.S. GDP in 1890 to 2.9% in 1998. (Again, see Odlyzko (2000) for these and many other fascinating statistics about communication.) Innovations such as mobile telephony and the Internet promise to continue the remarkable growth of this sector.

Industry	1994 Revenues (US$ billions)	1997 Revenues (US$ billions)	Annual Growth (%)
Telephone	199	256	9
Post	50	58	6
Advertising	152	188	7
Broadcast	106	126	6
Sporting Goods	54	64	6
Airlines	88	110	7
National Defence	282	271	-1

Table 1. Selected Sectors of the U.S. Economy

It is no surprise, then, that the communication sector has attracted the attention of policy-makers and economists. But there are further reasons, beyond just the size of the sector, that single out communications as an area of particular interest. These reasons relate to the characteristics of the networks that deliver communication services. The characteristics are not individually unique to communication networks, as the discussion will make clear. The combination of all four factors is, however, unique to networks. We argue in this paper that the combination represents a particular challenge to the study and regulation of communication networks.

The first characteristic is the *cost structure* of networks. Typically, there is a large fixed cost to building a network - digging up pavements to lay cables for a communication network, laying track or asphalt for rail or road travel, gathering all the information for the first edition of an encyclopaedia. For example, the estimated fixed cost of the British Telecom (BT) network in the U.K. accounts for approximately two-thirds of the total cost of the network (Oftel, 1999). Moreover, these costs are often sunk, i.e., irrecoverable, even if production stops. In contrast, the marginal cost of operating the network is low. On a standard telephone network, the marginal cost of making a local telephone call is of the order of 0.2-0.4 pence per minute. The marginal cost of producing an encyclopaedia is the cost of writing and shipping a set of CDs - a couple of pounds. The economies of scale that result from this cost structure often lead to "natural monopoly"- a dominant firm that captures the entire market.

There are two observations to make at this point. First, this cost structure occurs in many, non-network industries. On its own, then, this factor does not raise any issues that are not familiar to any economist who has taken an introductory micro course. Secondly, there is often considerable debate about whether fixed costs are all that important in communication networks. In the past, it has been thought that the access part of telephone networks (the last segment of the network, from the last switch to the customer's premises, called the local loop) was such a significant cost that provision of access should be monopolized. Recently, this view has been brought into question. The next section considers this issue in some detail.

Suppose that you want to send an e-mail. For most of us, we would first access the Internet by calling an Internet Service Provider (ISP) over a telephone line. The average e-mail message is broken into around 20 pieces, or packets, by the sending computer. These packets are sent over a standard telephone line to the ISP, using a modem to convert the computer's digital information to the analog waves that telephone lines transmit. Each packet is transmitted to the nearest 'router': a special computer, dedicated to receiving and

forwarding packets, that is the Internet equivalent of a telephone switch. The router consults a database to decide where to send the packet; it passes each packet onto another router, or to the destination if it is close enough. Once all the packets arrive at their destination, they are reassembled into the original e-mail and read.

This seems pretty involved just for a message like "I'll be home in half an hour". Of course, each user does not care about which routers have handled the individual packets in an e-mail. What matters is the joint function of all of the components that go in to sending an e-mail. In the language of economics, the different parts of the network are called *complements* - items that are worth more together than separately. This is the second factor that distinguishes networks.

Again, complementarity is not unique to networks. I do not value particularly the individual components of the starter unit in my car; I care only about whether it starts when I turn the key. Taken with two other factors, however, complementarity is particularly marked in networks. First, positive externalities (about which, see below) mean that there are large gains to connecting two networks. Secondly, the cost structure means that some network segments are owned by a small number of firms, leading to a bottleneck problem. This is the case with the "last mile" of telephone wires running into houses, which typically is owned by an incumbent telephone company. Any telephone company offering services have to gain access to the last mile. The economic principles behind efficient pricing and use of bottlenecks is the theme of the next section.

The third factor is *externalities* - the fact that the benefit that users gain from a network depends on how much the network is used. In a communication network, there are benefits to each individual to being able to communicate with others; the more users are on the network, the greater the total benefit. For example, suppose that each individual gains a benefit of 1 from being able to communicate with any other individual; and suppose that there are N individuals on the network. Then the total value of the network is the number of pairings $N(N - 1)$, which is close to N^2 when N is large. This square relationship between the number of members of a network and the value of the network is known as Metcalfe's law. There are also indirect benefits associated with a large network. The more members of the network, the more likely it is that new services will be offered over it. (Think about the increase in the number and range of programs on television over the last fifty years.) In short, networks are more valuable if there are more people using them. See Katz and Shapiro (1985) and Farrell and Saloner (1985, 1986) for seminal analyses of positive network externalities. The externalities can also be negative: my utility from driving to London on the M3 decreases

with the number of other cars on the motorway; surfing the Web in the afternoon is slower than surfing in the evening, because of the number people who are trying to access the same lines and sites. These externalities are an important feature of networks; whether they give rise to market failures that require regulation is the subject of section 3 of this paper.

The fourth factor of interest in networks are the *social obligations* that are associated with them. In one way or another, networks are often viewed as providing essential services. It has been a concern to policy-makers that these necessary services be available to all, regardless of their income or the cost of provision. For example, in the U.S., local telecoms operators are required to offer subsidies on basic services to low-income consumers, and to make services "affordable" in high-cost areas. In the U.K., BT is restricted to charging geographically uniform prices to ensure that high-cost (e.g., rural) areas are serviced. Even in developed countries, where most consumers are able to afford access to public switched telephone networks (PSTNs), universal service obligations (USOs) continue to be a sensitive political issue - all the more when applied to new services such as Internet access. In less developed countries, USOs are of central importance in the growth of PSTNs. But the imposition of USOs necessarily creates distortions in the pricing of network services. Section 4 examines the effect of these distortions.

2. Network complements and pricing

It has been argued in the past that network industries would exhibit strong economies of scale. For this reason, it was preferred to have a single supplier in order not to induce wasteful duplication of resources. In order to avoid excessive monopoly charges, the "natural" monopolist was either owned directly by the State or heavily regulated. This view has been challenged in the last two decades. Technological changes have drastically reduced unit costs in most network segments, and arguably the relevance of economies of scale is exhausted rather rapidly for low levels of output.[1] Poor quality of services offered by incumbents and the asymmetries of information between the regulator and the regulated firm offered additional arguments to liberalise the entire sector, even in the presence of economies of scale.

In the current scenario, intervention aims at being less intrusive, letting the competitive process deliver the intended outcomes. The regulatory attention is now devoted to the design

[1] Mobile communications, and more in general cell-based wireless communications exhibit more or less constant returns to scale. Mobile services are supplied using a grid of cells connected to the fixed network. As new areas are covered, or as traffic increases within a given area, the only option left to an operator is to invest in additional cells or to do "cell splitting". Also transmission (long-distance) is more or less subject to constant returns to scale globally, although locally it exhibits increasing returns (dark fibre is usually installed in excess).

of an appropriate market structure (i.e., whether or not to allow an incumbent to be integrated across various segments or to require separation of the various activities, the study of the entry process) and to the monitoring of particular behaviours of the incumbent. In particular, there is now less emphasis on final prices control while interconnection rules are subject to close scrutiny. The change in the practice of regulation has produced a similar shift in the literature. The new environment has lead to a large volume of theoretical work that makes clear the key role of access pricing as a Trojan horse that can preserve the efficient large-scale use of facilities while still letting all firms use the existing infrastructure after paying access charges to the firm that owns and maintains it.

Access pricing is crucial in all network industries, in communications, energy and transport, since they all involve various bottlenecks such as low-voltage distribution networks, local loops, airport gates, rail tracks. In many instances, inter-connection disputes arise in a context in which the access provider is the historic operator, which is active in all traditional markets – i.e., vertically integrated. The fact that the bottleneck owner is allowed to compete against other firms means that there is a danger that the incumbent will set access charges which make entry difficult or even deny access on reasonable terms. This may suggest that the access price should be set low, in order to contrast the anti-competitive attitude of the incumbent. However, if the access price is set too low, inefficient entry may occur. Moreover, if fixed costs are involved in the bottleneck, the regulator should ask how much the entrants should contribute to repay the fixed cost of a service that they use in order to supply their customers.

In the remainder of this section we review various contributions and feasible practical solutions. We first present an analysis that concentrates on efficiency. This includes efficient allocations (the best product mix for society for a given level of resources scarcity) and efficient production (the cheapest cost, for a given output mix). This analysis is essentially static, in that it takes as given the essential infrastructure (the "upstream" bottleneck) and concentrates mainly on the downstream market subject to competition. We show in sections 2.1-2.4 how the presence of an integrated incumbent makes regulatory intervention almost a necessity. In section 2.5 we consider what happens when the incumbent can also be challenged by potential entry in the bottleneck. We show there that the incumbent may want to use particular pricing policies (e.g., bundling) to deter entry in all segments, especially when they are complements. In sections 2.6 and 2. we consider what could happen once facility-based competition is sufficiently developed such that each network needs some other network to terminate its off-net calls. We argue there that the regulatory concern of

foreclosure is substantially reduced. This analysis leaves out an important question on how investments in infrastructure occur. In particular, expectations of the level of access pricing over time will generate a range of possibly interdependent investment decisions taken by firms. Much less is known on the dynamic properties of entry, investments and access pricing regimes. This aspect of the process, which has not been extensively analysed by theorists so far, is the subject of section 2.8.

2.1 Long run incremental costs

Long Run Incremental Cost (LRIC) is a long run measure of costs, hence it does not tend to overestimate the value of asset if one adopted historic cost accounting (or underestimate the cost in case of labour-intensive assets). Capital is included in its measure (depreciation is rightly considered as an economic costs), allowing potentially for full recovery. LRIC is a measure of the true economic cost of an asset and sends the right make-or-buy signal to alternative suppliers of infrastructure. LRIC should represent the long run equilibrium level of charges, hence it guarantees to achieve allocative efficiency. Although the details are quite complicated, the underlying principles for its computation are the following:

- Assets are valued and depreciated on a current cost account basis, giving the current replacement cost of a modern efficient asset;
- Operating capital costs are grouped together according to the cost type and cost driver;
- Cost-volume relationships are estimated showing how these costs change over the long run with volumes of the relevant cost-driver;
- Increments are defined. The cost-volume relationships then show the cost saving if an increment is no longer provided
- The total of these costs is the LRIC of the service increment.

While the concept of LRIC relies on a respectable theoretical background, the practice of the computation of LRIC is much more problematic. Firstly, LRIC is at odds with traditional depreciation practices (typically straight-line). In the presence of technological progress, straight-line schedules would underestimate the true economic annual cost. This practice may be sponsored by regulators that, by choosing slow depreciation schedules, can obtain lower current prices, an appealing feature when they want to encourage entry. However, this does not have any economic justification and it is also sustainable only if the regulator could

promise at the same time higher future prices, otherwise investments would never happen. Secondly, the computation of LRIC is simple for a single activity which involves fixed costs (it simply corresponds to average total cost), but it is much more complicated when an activity is an input to the production of two or more output (think of exchange switches). The definition of increments turns out to be crucial and it is often dictated by objectives other than efficiency. Thirdly, LRIC computations still involve a lot of discretion, in particular in the definition of a sensible cost of capital.[2]

2.2 Cost-based rules

Once the LRIC of a bottleneck is known, the next question to ask is what the access price should be. Imagine the following stylised situation. In order to provide one unit of final good, downstream firms need one unit of the upstream input that is produced by the bottleneck owner at a unit LRIC c_0 in change of a unit access charge denoted by a.

If all firms in the downstream sector are similar (in terms of technology and products), downstream firms undercut each other until price competition drives to zero all extra profits. The price charged to final users ends up equal to the marginal cost of each firm, which amounts to the sum of the access charge and any other cost incurred in order to transform the intermediate good. If we denote the latter by c, the final price would be $p = a + c$.

Without any other source of distortion, the best that could be done is to follow a *marginal rule*: the price to the final user (the consumer willingness-to-pay) should be set equal to the total marginal cost of production. The access price should thus be set equal to the marginal cost of production ($a = c_0$) and in the end the consumer price would be $p = c_0 + c$.

A marginal rule of this type is relatively easy to understand and to implement. This explains - at least in part - why forward-looking LRICs have been recommended for a liberalised telecommunication market by the European Commission (98/195/EC). But when is it appropriate? The answer is: only when there are no other distortions in the industry. On the other hand, distortions in the incumbent's retail prices may exist for various reasons, for instance because the bottleneck involves also some unapportioned fixed costs, hence a marginal rule would not allow to recover them. Another type of distortion arises when the incumbent's prices do not reflect its cost structure, because the incumbent is constrained by some social obligations to charge identical prices in different geographic region. In these

[2] LRIC can be derived using bottom-up and top-down approaches. Bottom-up estimates (based on engineering estimates) are more precise in enabling cost causation (since they are based on explicit parameters) and easy to

circumstances, a marginal rule is *not* the correct benchmark unless additional instruments are used simultaneously to relieve the access charge from additional tasks (more on this in the section on USO). Access charges purely based on LRIC are an appropriate benchmark when retail-level distortions are eliminated (for instance by tariff rebalancing) or dealt with using other instruments. On the other hand, the common practice to apply uniform mark ups to the LRIC estimates to recover unapportioned costs does not reflect much economic analysis (see also section 2.1 of Cave and Prosperetti in this issue).

2.3 Efficient component pricing rule (Retail minus)

The access pricing problem may arise in a context in which the regulation of access is separated from users' prices. Supposing that the final product prices are already fixed, then access price has no effect on allocative efficiency. The regulator may still be concerned with productive efficiency, that is to say with efficient entry and cost minimisation. As can be inferred from the analysis above, the downside of adopting LRIC access charges in such an unsuitable circumstance is that inefficient entry would be encouraged. Another option is then to move away from a "cost plus" approach based on estimated network costs to one in which access charges are derived from retail prices. The pricing policy that concentrates *only* on productive efficiency is the popular and controversial ECPR (Efficient Component Pricing Rule) also known as the Baumol-Willig rule, the imputation rule, the margin rule, or the parity-pricing formula.[3] The rule states that when final products are homogeneous and the market is contestable, the access charge should be equal to the difference between the final price and the incumbent's marginal cost on the competitive segment (c_1):

(1) $$a = p - c_1 = c_0 + (p - c_0 - c_1)$$

ECPR can be read in many equivalent ways:

- As a margin rule, it says that the margin of the incumbent in the final market (p - a) should be equal to its marginal cost in the downstream activity (c_1).

review. Top-down approaches are based on existing cost structures reported in the accounts. They are useful since they can reflect complex networks, however they are also more opaque and may hide inefficiencies.

[3] The rule was originally introduced by Willig (1979) and Baumol (1983). See Armstrong (2001a) for an excellent review of the literature on access pricing.

- As a parity principle, the bottleneck owner imputes itself for the bottleneck input the same price at which entrants buy the input, hence any attempt to practice a price squeeze would be detected by simple accounting separation.
- Productive efficiency is ensured. A potential entrant enters only if it is viable, which occurs only if firms are more efficient than the incumbent in the downstream activity.
- Entry does not alter the bottleneck cost recovery (revenue neutrality).
- Alternatively, the rule says that the access charge should be equal to the direct cost of providing cost (c_0) + the opportunity cost of providing access ($p - c_0 - c_1$) since this is the reduction in the incumbent's profit caused by the provision of access. In words:

(2) $a = \textit{direct cost} + \textit{opportunity cost}$

The simplicity of the formula explains in part its popularity. Revenue neutrality for the incumbent, on the other hand, is also the criticism made by opponents: *if* the incumbent is earning supernormal profits, they will continue to be earned also in presence of potential entrants. In this respect, the rule guarantees monopoly rents! However, the observation is not completely appropriate because ECPR *assumes* that final prices are optimally set.

The simplicity of ECPR is only apparent, since it derives from strong assumptions. In particular, entrants may offer new product varieties, so that consumer choice increases. Entrants may also be able to supply the bottleneck themselves, though using less efficient technologies. As shown by Armstrong (2001a), eq. (1) would have to be modified into the following more complex formula when demand-side and supply-side substitution possibilities are taken into account:

(3) $a = c_0 + \sigma(p - c_0 - c_1)$

In eq. (3) the opportunity cost to the incumbent is multiplied by a factor σ called "displacement ratio". The displacement ratio determines how much sales the incumbent firm loses as a result of supplying access to its rivals. The displacement ratio is generally less than 1, according to the degree of product differentiation, bypass opportunities and technological substitution. The opportunity cost of supplying access to rivals is typically reduced because there is not a one-for-one displacement of the incumbent's sales.

ECPR in its more complex formulation has never been put in practice. When advocated, it is generally in its simplest form given by eq. (1). In the UK, ECPR was discarded by the regulator in the mid-90's when a change was proposed in the computation of BT interconnection charges (Valletti, 1999). Perhaps to avoid the embarrassment of reintroducing the same concept after some time, the regulator has recently advocated its use under a different name - "Retail minus" - when it discussed the roaming charges that entrants would have to pay in the market for 3G mobile services for roaming on incumbents' networks in case negotiations failed. It is a bit ironic that ECPR did not find applications in fixed telephony at the time when incumbent's prices were regulated - in line with one of the assumptions of the proponents - while it is now proposed in mobile telephony where final prices are unregulated.

2.4 Ramsey charges and global caps

A final option is to recover fixed costs and common costs in direct proportion to service-specific incremental costs. This is the benchmark situation that emerges with a "benevolent" regulator quite familiar with the cost structure of service providers as well as their effort levels to minimise costs. This regulator fixes *all the prices* in order to maximise an unweighted sum of consumer well-being and total industry profits, subject to a break-even constraint for the incumbent. The optimal theoretical access charge can be rewritten as:

$$(4) \qquad a = c_0 + \sigma(p - c_0 - c_1) + \textit{Ramsey term}$$

in other words, it is very close to the generalised version of ECPR as given by eq. (3), plus an additional term that is related to the inverse of the elasticity of final users. Such extra term can be understood by noting that, in order to reduce distortions, customers of services that are not price sensitive should contribute more to the recovery of fixed costs. For example, if demand for calls from fixed to mobile users is less elastic than for long distance calls, then the access service of call origination should have a higher mark-up when requested by a mobile operator to terminate mobile calls, than when sold to an operator to provide long-distance calls. Downstream firms are like middlemen between the bottleneck monopolist and the final users. The access charge to the firm selling to users with rigid demand should therefore be higher than the access charge paid by another firm selling to consumers that are more price-sensitive. The formula is saying that optimal charges derive *both* from demand and supply analysis.

It is important to keep in mind that access is priced above marginal cost *not* because the incumbent exerts monopoly power but because deficits are socially costly and the charge performs as a tax used to raise money that repays the deficit. The charge is particularly high when it does not distort too much the allocation in the final market (the elasticity of the entrant's customers is low) or when the budget balance is particularly severe (equivalently, the social cost of public funds is high). By increasing *a* beyond its marginal cost of production, some retail prices can be reduced, which is in the interest of final users.

Ramsey charges imply that services that use the *same* bottleneck may end up paying *different* mark ups if their elasticity is different. However in practice many regulators would not allow this for the fear that the incumbent could engage in anti-competitive practices. For instance it is very likely that the EC would interpret Ramsey access charges as unfair and discriminatory. Such a fear is not always reasonable. As long as the regulator can ensure that market power is not abused, economic theory calls for different charges whenever demand elasticities are different, allowing for price reductions in more price responsive segments

Even more fundamentally, in order to implement Ramsey prices, a great deal of information is required. The regulator should know the cost of the regulated firm and also the different elasticities of demand. This kind of information is more likely to be in the regulated firm's hands rather than the regulator's. An immediate implication is that it would be better to delegate pricing decision to the firm, for instance using a *global price cap*, as proposed by Laffont and Tirole (2000), on the entire incumbent's range of products, treating the bottleneck input as a final good and including it in the computation of the price cap. The good properties of price cap mechanisms are well known and put in practice for the control of final prices. However, most regulators have resisted so far the introduction of global caps, preferring to control separate baskets for final and wholesale prices (this is for instance the way Oftel regulates in the UK).

To conclude the discussion contained in section 2.1-2.4, it is important to understand that different goals and policy objectives lead to alternative ways of calculating optimal charges. There is also one important practical corollary: the access charge is often performing too many tasks. While it is true that theory is useful to understand the mediating function of access prices, we stress that one first fundamental step should precede any access distortion: whenever possible, the use of access pricing as an instrument for the promotion of too many goals should be resisted and other instruments should be used. For instance, if the regulator believes there are barriers to entry, the tax/subsidy issue of the entry barrier should be addressed directly and be made explicit, rather than burying it into the access pricing problem.

The latter could indeed be the only option available, but only after having realised that other options are not feasible. In other words, by understanding the links between different problems, new instruments become available that allow fine-tuning of the regulatory process.

2.5 Entry and bundling

Arguably the biggest concern calling for regulation of access charges in network industries is that an integrated incumbent may use its monopoly position in some segments to sustain or extend market power in other segments that are potentially subject to competition. This is a classic problem of "leverage" that has attracted considerable attention in the literature on bundling. The analogy is close since an incumbent may make entry in a market unprofitable in different ways, e.g tying by product design, or "virtual" tying through pricing.[4]

In its simplest form the leverage theory has been largely discredited by the Chicago school. Imagine there are two markets, A and B. Market A is monopolised by an incumbent while market B is competitive. Consumers have a gross benefit V_A for product A and V_B for product B, *independent* from each other. In this scenario there is only "one" monopoly power to be exerted and the incumbent cannot gain any advantage if it tried to sell product A as a part of a bundle with product B. In fact, B would always be available at cost ($p_B = c$), hence customers would buy the bundle $A + B$ at the price p_{A+B} only if $V_A + V_B - p_{A+B} \geq V_B - p_B$, which is the same condition as $V_A + c \geq p_{A+B}$. In other words the bundle could just give the incumbent the same margin $p_{A+B} - 2c = V_A - c$ as if it sold A alone at the monopoly price.

The Chicago critique then shows that it is not possible to leverage monopoly power, hence pricing or design practices such as bundling and tying should be allowed *prima facie* since they are likely to be motivated by efficiency reasons such as cost savings.[5]

While the efficiency presumption is a serious starting point for the analysis of bundling, a closer scrutiny gives less reassuring results, in particular in the presence of risky investments, complementary products and network externalities. As demonstrated by Whinston (1990), the Chicago critique applies as long as market B is perfectly competitive. In the presence of imperfect competition, tying allows the monopolist in market A to commit to a more

[4] In the previous sections, a consumer would pay a price p to the entrant who, in turn, would pay an access charge a to the incumbent. The same situation could be described in an alternative way, with the consumer paying $p_E = p - a$ to the entrant for the "downstream" segment and $p_I = a$ to the incumbent for the "upstream" segment. Hence foreclosure can also be seen in two equivalent ways: the monopolist could either set an extremely high access charge in first case, or sell the joint product (upstream + downstream) as a bundle in the second one.

aggressive strategy in market B, blocking entry by differentiated rivals. Key to get this result is the presence of economies of scale in market B. In fact, to find it worthwhile to enter, entrants need a sufficient scale of production. Tying by the incumbent becomes a device for committing to a low price once entry occurs, hence it can eliminate competition in market B. The decision to bundle is the difference between making entry profitable and not profitable. If entry occurs, then bundling and its associated aggressive pricing would hurt the incumbent, so bundling is credible only if the monopolist can commit to it, for instance via product design.[6]

Bundling also gains a new dimension when entry is possible in all components. Rather than leveraging monopoly power into other markets, bundling may help the incumbent preserve its original dominance in a primary market. Carlton and Waldman (1998) investigate this idea by building on Whinston (1990) and examine the role of intertemporal economies of scope (i.e., an entrant is more profitable in a second period if it is also present in the initial period). A monopolist is sole producer initially, while entry can occur later on also in the primary market. In the complementary market there is potential for immediate entry. Tying can keep rivals out in the complementary market in the initial period, in order to make entry unattractive in subsequent periods in the primary market. This is because fixed cost cannot be recovered by operating in the complementary market in one period only.

Entry can also be deterred by discouraging rivals from investing in innovative activities. Bundling (or – equivalently - denial of access to an essential facility) can then be a formidable tool in the hands of the incumbent in risky and dynamic industries. To see this, imagine firm M is the incumbent for two perfectly complementary products, A and B. There are n identical consumers demanding one unit of a final product for prices that do not exceed their reservation value V. If it is not challenged by any rival, M can earn monopoly profits $\pi_M = n(V - 2c)$ in the overall market, where c is the unit cost of production for each component. There are 2 potential entrants, E_A and E_B, one in each segment. Suppose entrants have to make an up-front investment and, in case the investment succeeds, they become more efficient than the incumbent in their segment. Let denote by S the surplus resulting from each innovation (i.e., the unit cost saving Δ multiplied by the quantity produced $S = n\Delta$).

---

[5] Another reason to bundle would emerge even in uncontested monopolies as a form of price discrimination when there is a negative correlation between the reservation prices of goods A and B. Negative correlation is probably not too relevant in network industries with high product complementarity.

[6] The commitment problem disappears in Nalebuff (2000), where an incumbent sells complementary products. In this situation, the incumbent can achieve a better price coordination by bundling several components compared to individual component rivals (a well known result in IO that goes back to Cournot). This can be a stable equilibrium, in the sense that the incumbent prefers bundling to selling each component individually and,

Imagine first only one entrant is successful. There are many equilibria, where the surplus created by the innovation is split in different ways between the incumbent and the entrant. For instance the entrant may charge c for its own component, while the incumbent sets $V - c$. At the opposite end of the spectrum, the incumbent may charge $V - c + \Delta$, while the entrant is just left with a price that just covers its cost $c - \Delta$. Without changing the nature of the argument, we concentrate on the intermediate case and assume that the total surplus S is split 50:50 between the two parties.

If both entrants are successful, the incumbent does not sell anything. Also in this case there is a multiplicity of equilibria that distribute differently the total surplus $2S$. As before, we concentrate on the symmetric case. This means that each component is sold at the incumbent's cost c and each entrant obtains the full reward S for its innovation.

The outcome of entrants' investments is uncertain. The more an entrant invests, the higher the probability that the investment is successful. In particular, in order to get a probability p of a successful innovation, suppose a firm has to invest an amount $p^2/2$. We impose the restriction $0 < S < 1$ in order for the probability not to exceed 1, as will become apparent below.

We are now in a position to analyse a simple 3-stage game where the incumbent decides first whether to bundle the two components or not, followed by entrants' investment decisions and, finally, by price competition. Solving backwards, the solution to the last stage has already been described above. In the second stage, we have to analyse two possible cases. In the first case, there is no tying. This means that an entrant - if successful - can either sell its product with the incumbent if the other rival has not been successful, or with the rival if it has succeeded as well. The investment choice for firm E_A comes from the following problem:

$$(5) \qquad \max_{p_A} \pi_A = p_A(1 - p_B)S/2 + p_A p_B S - p_A^2/2$$

The solution at a symmetric equilibrium ($p_A = p_B$) is easily obtained where each entrant selects the following probability of success (the superscript n stand for the no-tying case): $p^n = S/(2 - S)$, which is a number comprised between 0 and 1 given the restriction on S.

On the other hand, in case the incumbent decides to tie its components, the entrant can sell if and only if both entrants are successful, hence it maximises the following expression:

---

at the same time, rivals prefers not form a rival bundle to avoid ruinous competition of bundle against bundle. This result is relevant for markets that produce "systems": Microsoft Office is the case in point.

(6) $$\max_{p_A} \pi_A = p_A p_B S - p_A^2 / 2$$

The first-order condition at a symmetric equilibrium is $p_A(S - 1) < 0$, implying that, with tying (superscript t), no entrant has any incentive to invest: $p^t = 0$.

It is immediate to see what are the effects of tying: it gives less options to the entrants, decreasing their incentives to innovate and to enter in the first place. Our example is extreme, but it keeps the flavour of the argument developed under more general conditions by Choi and Stefanadis (2001). The trade-off that the incumbent faces is also easy to recognise. On the one hand, if it ties it lowers the probability of joint entry that would completely displace the incumbent. On the other hand, without tying there is a probability that part of the surplus created by the entrant can be appropriated by the incumbent. In analytical terms, the incumbent's profit under tying and no tying are respectively:

(7) $$\pi_M^t = (1 - p^{t^2})\pi_M = \pi_M$$
$$\pi_M^n = (1 - p^{n^2})\pi_M + 2p_n(1 - p^n)S/2 = 2(1-S)(S^2 + 2\pi_M)/(2-S)^2$$

By comparing the previous expressions, it turns out that tying would be profitable only when $\pi_M > 2(1-S)$, that is when there is a big interest to protect monopoly profits. Notice that the previous inequality is particularly likely to hold when S is high. On the one hand, the incumbent could share part of the gains, however the rivals' incentives to invest would be very high, making it likely a joint success that would leave the incumbent with zero profits. Bundling becomes profitable when there is a high risk of being supplanted by low-cost entrants in both components.

In this example bundling is clearly inefficient. All consumer surplus is extracted and no cost-reducing investment occurs (notice that a social planner would invest even more than the no tying case since it would coordinate better than the two potential entrants). We should warn that this conclusion cannot be generalised. A full welfare analysis would have to take into account also the incumbent's incentive to invest.[7]

[7] The probabilistic nature of investments in system markets is also considered by Farrell and Katz (2000), but in a different model where an incumbent M is always a monopolist in one component, while there is R&D in a secondary market that can improve the final product. They show that M's integration increases its incentives to innovate while rivals would react by spending less on R&D. Integration brings some benefits, since it reduces double mark ups and, more interestingly, once investment has occurred, M has incentive to encourage efficient

2.6 Two-way access pricing: the call termination problem

Most communications involve two-way networks: calls initiated by a subscriber of a certain network may be terminated on a different network and, conversely, a certain network will terminate calls originated on other networks. There are revenues associated with terminating calls, and this should have an impact on outgoing charges that operators set to attract customers in the first place. Termination charges also feed directly into call charges when calls are destined to a rival network, making the problem even more challenging. In this section we neglect this additional effect - addressed in section 2.8 - and consider a simpler situation where termination revenues arise from a separate market. A relevant example is the so called fixed-to-mobile termination problem.

It is a common practice in mobile telephony that the party that makes and pays for the call is not the same as the party that chooses which operator will terminate the call. This system, known as CPP (calling party pays) is adopted almost in every country in mobile telephony, with the notable exceptions of Canada and the US where there is a RPP system (receiving party pays). Under CPP, there is a striking discrepancy between the interconnection rates for fixed-to-mobile services compared to the interconnection rates paid for mobile-to-fixed services.[8]

Once a person has decided to join a particular mobile operator, that operator has a monopoly position over termination services to its subscribers. The decision to subscribe to a network has an influence on the price charged to all other customers that may want to call that person. Hence termination services involve an externality problem that is a potential source of distortions.[9] It should be noted that the termination problem is not peculiar to mobile

---

entrants since this allows *M* to practice a better squeeze (*M* can offer consumers as much surplus as possible in the complementary product in order to extract high surplus from primary market). *M* does not engage in the squeeze to earn more money on that product, rather it lets the competitor sell but extracts profits in the complementary market. Integration strengthens the squeeze since *M* has an additional instrument to force low prices, either via pricing or via producing a better product; however the downside is that *ex ante* incentives to invest are reduced.

[8] According to OECD (2000) the ratio of average interconnection rate for fixed-to-mobile was 11:1 compared to mobile-to fixed in OECD countries adopting CPP. While there are difference in costs in the two services, they are hardly enough to justify such a difference. To confirm this, the same ratio was 1:1 in the US.

[9] A second source of distortions is related to consumer ignorance. In its inquiry into mobile termination rates, the UK Monopolies and Mergers Commission found that fixed-line users had little knowledge of the mobile network they were calling and of the call price (MMC, 1998). If fixed-line users base their calling decisions only on an estimated price based on mobile market shares, then the link between a specific termination charge set by a network and the number of calls terminated on that network is broken. If a mobile network raises its termination charge, it gets the full benefit and shares with other mobile networks the reduction in the number of calls received. As a consequence, networks will have an incentive to set very high termination rates. This problem is

telephony but is common to all network operators. In the context of mobile telephony, the subscriber base of fixed users is large, hence the number of calls potentially terminated on mobile networks represents an important source of revenues for mobile operators.

In a stylised situation where the mobile sector is perfectly competitive and mobile operators charge two-part tariffs to customers with identical preferences (for instance a monthly fee and a charge per minute for every call made), then operators would compete to attract customers by setting each call charge equal to its marginal cost and then set the fixed component to divide the surplus created between the operator and its customers. If, as assumed, the mobile industry is perfectly competitive, operators would earn zero extra-profits. Any increase in termination profits (for instance because the termination charge is set above its cost) would simply be passed to mobile subscribers via lower fixed charges. Fixed charges may even become negative, as long as considerable extra-profits arise from termination: this result may explain handset subsidies, a common feature in many mobile markets.[10]

It is clear that, even with perfect competition for mobile users, there is little competition for providing access to mobile subscribers. This remark suggests that if mobile operators are free to determine termination rates, they will set charges that extract all possible surplus from fixed users. In principle then some regulatory intervention can be beneficial.

Welfare considerations on termination rates are complicated, since an increase in termination charges both increases fixed-to-mobile calls and decreases fixed fees to mobile users. As discussed by Armstrong (2001a), marginal cost pricing (implying no subsidies for mobile connections) is the correct benchmark when a certain number of stringent assumptions are satisfied. In particular, the demand of mobile subscribers should be rigid with respect to subscription decisions, there should be no monopoly power exercised by the fixed network and there should be no network or call externality. If mark-ups over termination charges are added by fixed operators, they should be counteracted by setting termination charges below cost. On the other hand, above-cost charges would be beneficial in the presence of network externalities, since higher termination revenues could be used to subsidise entry, thereby raising the equilibrium number of subscribers that benefits everybody, because both fixed-line

---

potentially exacerbated by the adoption of mobile number portability since there could be no correspondence between a certain number and the current network choice of a subscriber. Carrier identification should then be promoted in order to make termination services more competitive.

[10] Handset subsidies and, more in general, subscription discounts, are also related to the presence of switching costs (Klemperer, 1995). With consumer switching costs, a firm is typically willing to serve to a larger set of customers in the first periods than in traditional models because this enlarges its "captive segment" of the market

and mobile customers are able to call and be called by additional subscribers. As a result, unregulated above-cost termination charges may be "good" in the initial phases of mobile development since they increase cellular penetration rates.

The potential market failure associated with termination services may be considerably diluted if people care about receiving calls. In our discussion, we have implicitly assumed that users either do not receive calls, or do not place any value on incoming calls. On the other hand, it is more plausible that mobile phones are purchased with a desire to receive calls as well as to make them. If a mobile user places similar weights to calls made and received, then any attempt of a mobile operator to set high termination charges would induce subscribers to change network, since they would otherwise receive too few calls. Notice that this result is true even if the caller and the receiver do not belong to the same "closed group". The argument that people may care about receiving calls is particularly compelling in the mobile sector since a mobile phone gives a customer the ability to be reached by other people at any time in any place. Unfortunately there is not enough econometric evidence on calling patterns in mobile telecommunications to say a final word on the relevance of the termination problem and this is an area where further research is needed.

To conclude, it has to be recognised that - under CPP - there is a potential market failure, independently from the share an operator might have. There are many ways to cure it. Termination charges based on LRIC is an option that, as we argued before, is legitimate under some circumstances. However this kind of intervention would be rather heavy handed. An alternative would be to try to put a downward pressure on termination charges, using a price cap over the entire bundle of services offered by a mobile company. This is also quite an interventionist approach that many operators would want to avoid. In a related vein, the Australian regulator ACCC has recently decided that any discount that mobile operators offer to their customers would have to be passed also on termination. The obvious downside to this is that, in the anticipation of the additional effect, operators would be more reluctant to fight against each other to attract mobile customers.

Since the fixed-to-mobile termination problem has not arisen in North America under RPP, perhaps this is an alternative pricing arrangement that regulators might want to consider.

---

in the following periods. In the specific context of mobile communications, the presence of switching costs is also important to understand the vertical links between operators and service providers (Valletti, 2000).

However, growth rates have been slower in North America compared to Europe.[11] Given the current level of penetration, this makes RPP arguably an interesting option for 2G mobile telephony since there is no further need to subsidise the subscriber base. However, also in this case there is a downside that may produce unintended effects. Recall that the intervention is lead by the desire to pass some more benefits on to fixed subscribers. Under a RPP, it may happen that - despite the decrease in the price of fixed-to-mobile calls would increase attempted calls - the actual number of completed calls is diminished since the receivers would keep their handset switched off more often than under CPP. Hence fixed users may be worse off under RPP.

To strike the right balance among these alternative options is a delicate and challenging regulatory task. By going into the very nature of the problem in question one may find better solutions that do not bring potential costs from regulatory failure. For instance, the call termination problem exists because operators do not compete *directly* over this type of services. The best solution would not be to intervene by setting the charges, but rather take steps to eliminate the bottleneck. For instance, giving the customers the opportunity to choose two operators - one for origination and one for termination of calls - would create direct competition in both markets, without having to worry too much about the linkages we discussed above.[12] Clearly, there would be some costs associated to this unbundling proposal, for instance the handset would have to contain two SIM cards, however this is the kind of creative solution we hope that regulators will be looking for in the future, bringing benefits without being intrusive.

2.7 Two-way access pricing and competition: open issues

The call termination problem described in the previous section is rather extreme in the sense that it is relevant when the market of callers is completely separate from the market of receivers (e.g., mobile users and fixed subscribers in the previous section). Under fully fledged competition, operators would be competing for the same customer base that would *both* originate and terminate calls (this would happen under convergence of fixed and mobile telephony). As long as operators *A* and *B* both command some market share, operator *A* needs

[11] Many other factors also play a role, most notably in the US there are several competing and incompatible digital standards. See Gruber and Valletti (2001) for a survey on mobile telecommunications.

[12] Under CPP the customer would still be financially responsible only for outgoing calls and hence will choose the operator that offers the cheapest package for originating calls close to the customer's profile. The customer

interconnection with *B* to terminate the calls that *A*'s customers destine to *B*'s customers and vice versa. There is a sort of "double coincidence of wants" that makes the interconnection problem less problematic since the foreclosure problem typical of one-way access would seem to disappear. In a symmetric situation access charges may even be thought to be irrelevant since *A* pays *B* the same amount it receives from *B*. This intuition is correct to the extent to which foreclosure is not a great danger when operators have both developed their customer bases and they have an incentive to conclude successfully commercial negotiations over the interconnection terms. However, this does not imply that regulation is not needed any more in such an environment.

A first concern arises when access prices can be used as an instrument of tacit collusion (Armstrong, 1998; Laffont *et al*., 1998). Collusive (i.e., monopoly) prices can be sustained using high access charges because of a raise-each-other's cost effect. To see this, imagine what happens if operators charge monopoly prices to customers. If customers call each other with the same probability, the traffic is balanced and an operator pays the rival for termination services the same amount it receives from the rival for similar services, independently from the value taken by the access charge. This can be an equilibrium only if no one has a unilateral incentive to deviate. If one firm deviates from the monopoly charges by undercutting the rival, it induces its subscribers to call more. Since part of the calls made are destined to the rival's network, the effect of a price cut is to send out more calls than it receives on–net from the rival. The resulting net outflow of calls has an associated access deficit that is particularly burdensome if the unit access charge is high. This will discourage underpricing in the first place. To get this result some conditions are needed, for instance products need to be not too homogeneous, otherwise the incentive to undercut would have the additional benefit to get market share.

Perhaps more crucially, another condition that is needed to generate this non-cooperative collusive result is that tariffs are linear. Once firms are allowed to offer non-linear prices the result collapses. For instance, with two-part pricing, it is still true that a high access charge feeds into high retail charges. However, all the profits generated are used to lower the fixed component (an example is call termination described in the previous section). Pricing itself may become efficient, since operators would tend to charge call prices equal to their perceived marginal costs: this result typically occurs when operators have more instruments to build

---

will also be inclined to choose separately the cheapest package offered by competing operators to terminate calls since he will anticipate that more people will be willing to call him.

market shares without having to inflate their outflow charges. As we have just described, these models also tend to generate profit neutrality with respect to access charges. This feature is an artefact of the symmetry considered by most models, and it typically disappears when firms differ in their cost structures, or when there is partial market participation or biased calling patterns among consumers (Dessein, 2000).

More sophisticated pricing policies are interesting since they better reflect the reality of pricing practices among operators. For example, mobile operators may price discriminate between calls destined on-net and off-net. In this case profits would depend on access charges. In particular, *low* access charges become profitable overall since customers become a liability and firms are less willing to fight aggressively for market share. While price discrimination may well be dictated by efficiency reasons reflecting customer heterogeneity, their downside is that they can also be used anti-competitively. For instance, if consumers care about incoming calls, off-net charges would tend to explode in order not to give rival's subscribers benefits from receiving calls. High access charges would cause a *de facto* connectivity breakdown (Jeon *et al*., 2001).

Another form of pricing structure that can be employed is to charge for incoming calls (RPP). Jeon *et al.* (2001) show that competitively determined charges would follow the "off-net-cost pricing" principle. An operator would set prices *as if* all subscribers belonged to the rival network (even if they are shared in equilibrium). For instance, if we denote by p the outgoing price, r is the incoming charge, a the access charge, c_t the termination cost and c_o the origination cost, then it results: $p = c_o + a$ and $r = c_t - a$. It is clear that a affects the level of traffic and the allocation of costs between the different sides of the market, so in principle regulation could be beneficial. Under RPP and price discrimination, as before, operators would internalise any externality on on-net calls. Still there would be strategic manipulations on off-net calls. Connectivity breakdowns could occur but for a different reason: receivers would hung up in order not to pay excessive charges for incoming calls. Notice that this setting is particularly relevant for an Internet environment where some customers (websites) mainly send information and the other side of the market (consumers) mainly receives information.

This discussion should have made clear that access charges under competition are still an area of ongoing research. The different models that we have reviewed reflect to some extent the existing patchwork of interconnection regimes based on different assumptions, historical distinctions and regulatory objectives. Such differences are becoming unsustainable in a world

of convergence, hence to preserve many different regimes may be extremely inefficient and encourage arbitrage opportunities. To avoid this, DeGraba (2000) has recently advocated a generalised "bill and keep" system, where a called party's carrier would not be able to charge an interconnecting carrier to terminate a call. When both parties – the sender and the receiver – benefit from a call, it is efficient that they both share the costs. This does not happen when operators can exercise monopoly over termination fees, which is a typical feature of CPP systems. DeGraba argues that competition works more effectively when operators recover their costs from end users who can choose among competing carriers. A bill and keep system would then reduce the termination monopoly access problem. Such a system may also resolve a second source of regulatory inefficiency, since termination charges are typically structured as per-minute charges, while most network costs are based on required capacity, allowing carriers to achieve efficient end-user rate structures. Finally, by eliminating inter-carrier termination charges, artificial cost differences between on-net and off-net calls would be eliminated. The Federal Communications Commission (FCC) is known to be examining bill and keep as a solution to termination charges in the US, with particular reference to ISP reciprocal compensations.

2.8 Access prices and investments

One of the most important issues in the economics of regulation is how to encourage firms to invest in infrastructure. Intuitively, there is a trade off between optimal access regulation in a static framework and in a dynamic one. If static regulation reduces the use of monopoly power over the infrastructure, then it also reduces profits that can be earned by the investor/owner of the facility. Access regulation based on simple cost recovery rules, while encouraging efficient utilisation of assets, may risk discouraging investments. The reason is simple. If operators rationally anticipate that, once somebody has invested, then the regulator will grant access at cost, everybody will then wait for the investment to be done by somebody else and then seek access.

This is a typical free rider problem that may cause big losses in social welfare. At best, investments are reduced; in the limit there may even be no production at all if no one invests in infrastructure. If this happens, it is easy to argue that there should be no access regulation since reduced competition is better that no services being supplied. This is clearly an extreme statement, but it should be taken seriously. In the presence of sunk costs, regulation of access terms and prices affects the return an infrastructure owner can expect to receive as a result of

its investment efforts. In economic terms, the nature of *ex post* access regulation has an impact on *ex ante* incentives to invest. Notice that, in the presence of sunk costs, the hold up problem is not just typical of regulatory appropriation, but may emerge in a similar fashion in an environment with commercial negotiations and contractual incompleteness.

In the presence of infrastructure competition and bottlenecks the regulator's problem becomes particularly challenging and involves many trade-offs. One is the desire to have a downstream level playing field while ensuring the incumbent to recover its upstream fixed costs or some social obligations. The regulator may also want to promote particular *entry modes*, where the typical dilemma is between facility-based and service-based competition. In telecommunications for instance, in the first case both the incumbent and the entrant build their own backbones and local loop facilities, so it may involve unnecessary duplication of infrastructure. This does not happen in the second case where the entrant leases the incumbent's access facilities; however the environment becomes much more intrusive, while in the first case the regulator can rely more on direct competition than on regulatory intervention. Since suitably adjusted access charges can encourage one particular entry mode, it is clear that investments will respond to access regulation by flowing at different network levels.[13] In addition, the regulator should be concerned about the timing and choice of investments. This issue is particularly important in industries with high technological progress. Since very little is know about this latter class of normative questions, it is helpful first to address the case of what would happen without any regulation at all. Infrastructure owners may want to maximise the use of their facility since its intensive use would reduce the average cost to all users. However, this desire clashes with another one, since the infrastructure owner would also try to reduce downstream competition, which implies a reduction of access to the infrastructure by its rivals.

Imagine a situation where a network has to be built and the investment cost declines over time due to technological progress. An incumbent operator first decides whether and when to invest. Then a rival chooses whether and when to seek access. Finally, if access is sought, the two parties bargain over the terms of access. As is standard in economics, this game has to be solved backwards. In other words, the investment choice is contingent on the expectations about the rival seeking access and on the outcome of negotiations. Imagine also that products are sufficiently differentiated so that the use of the investment is non-rival and infrastructure

[13] Cave and Prosperetti in this issue show the existence of a "service bias" in European telecommunications regulation, aimed mainly at opening incumbents' *existing* infrastructures. In particular, see their section 3.1

owners do not fear the rent dissipation caused by downstream competition and have an incentive to optimise the use of the facility.

In the last stage of the game, negotiations can only be over variables that can be altered at the time of negotiation. As the investment has already taken place, infrastructures themselves are sunk and cannot play a role during negotiations. This typically weakens the provider position. By denying the rival the use of the infrastructure, it gains nothing and loses whatever access charge it might receive. There is also another aspect that crucially affects the scenario. The access seeker can in fact become the provider itself and sell access to the rival. In this case negotiations would be reversed. There is a potential for both firms to "race" in order to be the first to provide the infrastructure. By doing so, an operator avoids the access payments and receives access revenues. This gives a reason to pre-empt rivals and incentives to invest are then raised. The race to become the "common carrier" speeds up the operators' choices. However it is not clear if timing choices are aligned with the social optimum. The racing process may go too far and investments happen too soon.[14]

Access issues become of greater concern when firms that use the infrastructure are also direct competitors of the infrastructure owner. If competition effects are extreme, the infrastructure owner will not grant access unless required to. Here regulation plays a crucial role. The entrant is obviously keen on obtaining access. Without compensation, however, the incumbent will wish to delay investments. This can be solved by requiring the entrant to bear more of the costs. But for the regulator this increase might reduce the possibility of entry itself. The regulator should try to manage this tension between investment incentives and timely competition. An access price régime can be used by the regulator to create competition between industry participants over the *provision* of facilities. If a firm "wins" in the provision of infrastructure, it becomes the common provider and receives access payments from other firms. If it loses, it will either pay for access or duplicate the infrastructure. By committing to an appropriate access rule, the regulator can directly determine the difference between winning and losing for operators.

The existing theoretical literature has not come up yet with a general answer to this intricate problem. The possible trade off between static and dynamic efficiency that we

---

where it is argued that such system had a negative impact on investments in access networks, and section 4.2 on Local Loop Unbundling.

[14] Gans (2001) shows how, in the simple context of a race between alternative suppliers of very differentiated (non substitute) goods, a simple rule that apportions capital costs according to the relative economic profit that is expected to accrue to the access provider and to the access seeker would allow them to achieve optimal investment choices.

highlighted at the beginning of this section should not be taken as the only possibility since regulation interacts with other important variables such as market structure and entry conditions, competitive behaviour of market participants, and technological progress. For instance, the unintended outcome of bad regulation could be to achieve low levels of both static and dynamic efficiency. This could be the situation in mobile telephony if too little spectrum is made available to a handful of companies that do not compete against each other and do not need to adopt innovative technologies if they are protected against entry by licence conditions. Conversely, under some circumstance it is possible to achieve the best of the possible worlds, i.e., high levels of both static and dynamic efficiency. In this situation operators would be competing against each other, achieving relatively efficient allocations, while still securing profits that create the incentive to invest. The presence of strong network externalities can support a case like this one.

3. Demand-side effects

The previous section considered the issue of pricing access to bottlenecks. Networks want access to each others' customers because of positive externalities - the fact that consumers' valuations of a network's service or product increase with the total consumption of the service or product. For example, the value from joining a mobile telephone network is higher when that network has a greater number of roaming agreements with other networks. In the first part of this section, we look more generally at the effect that positive externalities have on the market structure - the number of firms and the degree of competition - of network industries. We start by supposing that networks are incompatible: the positive externalities arising from a network depend only on that network's size. We argue that expectations play a crucial role in these industries. Many different market structures are possible, depending on expectations; some of these market structures are concentrated, involving a small number of firms and a low degree of competition.

The discussion then moves on to consider compatibility and interconnection between networks. What incentives do networks have to be compatible or to interconnect? This question is crucial for many networks, and particularly at the moment for the Internet. The Internet is all about connectivity: any two computers anywhere in the world can, in principle, communicate with each other. This can occur only through thousands of interconnection agreements between the many separately-owned communication networks that comprise the Internet. The functioning of the Internet therefore depends on connectivity arrangements.

In the last part of the section, we reverse matters and look at how networks should price when externalities are negative i.e., congestion is present. This is a question of great relevance at the moment for the Internet. Increasing use of the Internet for more and more demanding applications has lead to considerable congestion at certain parts of networks. Internet Service Providers (ISPs) are very much interested in pricing schemes that can offer Internet users the correct incentives so that congestion is reduced and revenues are generated to fund increases in network capacity.

3.1 Market structure and oligopoly pricing with positive externalities

You might have experienced a problem when speaking into a microphone when close to a speaker. A small noise going into the microphone is amplified and fed through the speaker, back into the microphone, increased further out through the speaker, and so on; the result is a very loud, unpleasant noise (unless you are Jimi Hendrix). This is an example of positive feedback - a small initial input amplified into a large final output. An analogous effect can occur in markets with positive externalities. When two incompatible networks are of similar size, consumers will have a slight preference (all other things being equal) to join the larger network. But then this network becomes even larger than the other network, and consumers are even more willing to join that network; and so on. An initial small difference between the networks is amplified into a large final difference.

The positive feedback mechanism has several implications. The first is that there are many possible market outcomes. If network A starts off with a small lead in market share over network B, the positive feedback amplifies that difference and A becomes dominant. On the other hand, if network B has the initial advantage, then B becomes dominant. This is a typical co-ordination game. Table 2 shows stylized payoffs for this type of game between two players who each must choose one of two actions: driving on the left- or right-hand side of the road, say. If both choose the same side, then all is well and they both receive a high payoff of 1. If they choose different sides, then a crash results and they both receive a low payoff of –1. There are two (Nash) equilibria of this game. If player 1 chooses ‘left’, the player 2’s best choice is also ‘left’; and vice versa. Hence one equilibrium outcome is for both to choose ‘left’. But clearly the same applies to ‘right’, and this too is an equilibrium of the game.

	Player 2	
	LEFT	RIGHT

Player 1	LEFT	1,1	-1,-1
	RIGHT	-1,-1	1,1

Table 2: A Co-ordination Game

The game in table 2 describes demand with positive externalities. Instead of 'left', read 'network A', and read 'network B' for 'right'. Then there are two equilibria: one in which both consumers choose to join network A, and one in which they choose to join network B.

The second implication of positive feedback is that expectations are all-important in determining how many networks can operate profitably. The network that is expected to be larger is more attractive to consumers, and so they are more willing to join that network. The original expectations are borne out. In terms of table 2, expectations determine which of the two equilibria occurs.

Thirdly, in order to complete the story, the supply side i.e., the behaviour of the networks, must be considered. Suppose that demand takes the simple form in table 2: each consumer demands one unit of the networks' products and has a gross surplus from consumption described by the payoffs in the table. Further, suppose that the networks' can produce at zero cost. In this case, equilibrium output and pricing is straightforward. If expectations are that network A sells to both consumers, then the equilibrium in which expectations are fulfilled involves network A producing two units and network B none. This is the Cournot-Nash equilibrium. (The price that network A receives in this equilibrium is indeterminate; but, as long as that price is between zero and 1, network A is willing to sell and the consumers are willing to buy.) In terms of the Bertrand equilibrium in prices, network A charges a price of 1 and receives a demand of two units, while network B receives no demand at any positive price.[15] Of course, equivalent output and price decisions hold in the other equilibrium.

More complicated demand structures lead to more complicated output and pricing outcomes. Two conclusions are common to most demand cases. First, holding the number of networks that operate in equilibrium fixed at greater than 1, positive externalities make competition more intense. (This effect is not observed in the simple example above, since only one network operates in equilibrium.) The reason is easiest to see when networks compete in prices, offering differentiated products. Without positive externalities, each network has an incentive to undercut its rival's price because by doing so, it can increase its demand. The greater the price elasticity of the network's demand, the greater the incentive to

[15] Consumers have the option not to buy from either network, in which case they receive a surplus of zero.

undercut. This same effect is present with positive externalities; but once a network's demand increases due a price cut, the network becomes more attractive to consumers, and so its demand increases by even more. This second round of effect means that demand increases more with positive externalities i.e., the price elasticity of demand is increased, and so competition is more intense. Although this story is told in terms of price competition, a similar intuition holds when networks compete in quantities. See Katz and Shapiro (1985).

Secondly, positive externalities can lead to concentrated market structures. This is (almost trivially) true of the game in table 2: in either equilibrium, one network has all consumers. To illustrate more general cases, consider a simple model (based on Katz and Shapiro (1985)) in which two networks can each produce zero, one or two units of a homogeneous good. Each network's inverse demand is given by $2 - Q + q^e$. Here, Q is the total output of the two networks (a number between 0 and 4) and q^e is the expected output of the network. For simplicity, there are no costs of production; and the networks take consumers' expectations as given when choosing their profit maximizing outputs. Consider two cases. In the first, it is expected that both networks produce one unit each. Suppose that one network produces one unit, and consider the best response of the other network. If it produces no units, it earns a zero profit; if it produces one unit, it earns a profit of 1; if it produces two units, it earns 0. Hence it should produce one unit. It is therefore an equilibrium, in which expectations are confirmed, for each network to produce one unit each. Now consider a second case in which it is expected that one network produces two units and the other zero. Suppose that the second network does produce nothing. The first network's profit is zero, 3 or 4 if it produces zero, one or two units. With the first network producing two units, the second network's best response is to produce nothing. Hence there is a second (Cournot-Nash) equilibrium in which one network dominates.

The discussion so far has assumed that networks are incompatible or not interconnected. The basic incentives to interconnect can be seen by looking at the simple model of network value that leads to Metcalfe's law. Recall that this model implies that the value of a network with N members is $N(N - 1)$. Suppose that there are two networks, one with N_1 members, the other with N_2 members. If the two networks do not interconnect, then their individual values are $V_1 = N_1(N_1 - 1)$ and $V_2 = N_2(N_2 - 1)$. If the two networks interconnect, then their values become $V_1' = N_1 (N_1 + N_2 - 1)$ and $V_2' = N_2 (N_1 + N_2 - 1)$. Network 1's value is therefore increased by $V_1' - V_1 = N_1 N_2$; network 2's value is increased by $V_2' - V_2 = N_1 N_2$. Here, then, are the networks' incentives to interconnect: both networks values are increased by the same amount, N_1N_2.

This simple story ignores competitive effects between the networks; this can either encourage or discourage compatibility, depending on the situation. In general, interconnection decreases competition between equally-sized networks. To see why, suppose that two networks offer horizontally differentiated products and compete in price; that they start by charging the same price; and that one of the networks considers cutting its price. When the networks do not interconnect, the price cut has two rounds of effect, as above. When the networks interconnect, the first round effect is unaltered: customers are attracted to the network that lowers its price. But the second round effect disappears: a network that is larger is no more attractive than a smaller network, since a consumer joining one network gets the full benefits of all consumers on all networks, due to interconnection. Hence the price elasticity of a network's demand is lower when networks are interconnected, and price competition is less intense. (See Farrell and Saloner (1992) for an analysis of compatibility and price competition with network effects, Matutes and Regibeau (1988), Economides (1989) and Einhorn (1992) without network effects, and Katz and Shapiro (1985) with quantity competition.) The general conclusion, then, is that equally-sized networks will wish to make their products compatible.

When networks are asymmetric, different conclusions emerge. Katz and Shapiro (1985) show that larger firms tend to be against compatibility, while smaller firms tend to be in favour of compatibility. The reason for this is that, in their model, full compatibility makes all firms symmetric in equilibrium: since the network benefit is the same regardless of which firm a customer buys from, there is no force to make firms different. With incompatibility, however, if one firm is expected to be larger than another (for whatever reason), then these expectations are borne out in equilibrium. The larger firm makes higher profit in the asymmetric equilibrium than in the symmetric case, and so opposes compatibility. See also Mason (1999) for a model in which interconnection may decrease networks' profits, and for a survey of incentives towards compatibility and interconnection.

The asymmetric situation describes better past and current interconnection situations. Standards wars - fights to establish which of several incompatible technologies will dominate - are common enough. Shapiro and Varian (1999) discuss many interesting historical examples: railroad gauges in the U.S. in the early nineteenth century; international postal systems; and the early days of the telephone. A modern-day example of an interconnection war is instant messaging (IM), a technology that allows computer users to detect when their friends are on-line and to type out real-time messages to them. The market leader for IM is America Online (AOL), whose AIM service has 21.5 million customers, compared to the 10.6

million customers of the next most popular service, Yahoo! Messenger (see http://news.zdnet.co.uk/story/0,,s2084246,00.html.) The tremendous success of IM (which is offered free to customers, but generates plentiful advertising revenues) encouraged Microsoft, Yahoo! and several other companies to launch their own IM clients, based on the protocol used by AIM and hence capable of interconnecting with the AIM service. AOL altered its protocol to lock them out, and repeated this action when Microsoft found a way past the block. AOL has declared that it will continue to take active measures to block other firms from interconnecting to its IM service.

A second example of current interconnection disputes is occurring in the Internet. In the early years of the Internet, networks operated a 'bill-and-keep' or peering system, in which no settlement payments were made. (See Srinagesh (1997) for a discussion of interconnection arrangements between packet-based networks making up the Internet.) Each network carried others' traffic without charge - the underlying assumptions being that either flows were roughly symmetric, or any other arrangement would stunt the growth of the Internet. The transition of the Internet from academic to commercial, large increases in traffic volumes, and the unequal development of networks have put this system under considerable stress.

In 1996, the extensive peering arrangements agreed under the Commercial Internet Exchange (CIX) started to dissolve. Large networks argued that they received little benefit, yet incurred substantial costs, from interconnection with small networks; this contrasted with the net benefits gained by the smaller networks from access to the customer base of the larger networks. Large networks began to apply pressure on smaller networks to change the relationship from peers to supplier-customer; instead of bill-and-keep, small networks would make settlement payments to larger networks. In 1997, UUNet, a large ISP, informed 15 smaller ISPs that their peering arrangements would be cancelled; this was followed by UUNet's withdrawal from the CIX. At the same time, MCI and BBN, two other large ISPs, left the CIX agreement, meaning that three out the four largest networks in the U.S. were no longer part of the CIX.[16]

The larger networks continue to interconnect between themselves on a peering basis. The gulf between large and small networks has widened progressively with the consolidation taking place in the ISP industry. By November 1997, it was estimated that the U.S.'s four largest networks (UUNet, MCI, BBN and Sprint) accounted for between 85 and 95 per cent of

[16] UUNet have responded to criticism about their policy by publishing guidelines stating when UUNet is prepared to interconnect with a smaller network; the guidelines are reported in OECD (1998). UUNet reserves

total backbone (i.e., core) Internet traffic, with the remaining volume carried by upwards of 40 other, small networks; see OECD (1998). There is a growing fear in the industry that large networks will use their size to limit competition in the ISP market by excluding smaller networks from interconnection agreements. See Crémer et al. (2000), who advised GTE on the Internet aspects of the WorldCom/MCI merger, and Cave and Mason (2001).

This issue continues to be of central important to policy-makers. Most policy positions are based on a suspicion that large networks have market power; and they exercise this market power through interconnection agreements with smaller. Despite considerable research on interconnection, there is no consensus among theorists, and much remains to be done before interconnection and foreclosure is understood.

In addition, more research needs to be done on the effect of positive externalities on market structure. The current result is (roughly speaking) that expectations determine equilibrium; and there are many equilibria that are consistent with various expectations. These facts are common to many economic settings with positive externalities, or complementarities. For example, in Matsuyama (1991), workers' productivity in the manufacturing sector of a two-sector model is higher, the greater the number of workers employed in the manufacturing sector. In Diamond and Dybvig (1983), investors must decide whether to withdraw or roll-over a deposit to a bank; the payoff from withdrawal (rolling-over) is increasing in the number of other investors who withdraw (roll-over). In all of the models, there can be multiple equilibria: if all agents expect one outcome (all to work in the manufacturing sector, join a network or roll-over), then it is optimal for each agent to act in the same way; if all expect another outcome (all to work in agriculture, not join a network or withdraw), then it is optimal for each agent to act in that way.

There has been much interest recently in ensuring uniqueness of equilibrium in these settings. Carlsson and van Damme (1993) argue that in *global games* (in which the actual payoffs to the game in table 2 are observed by the players with some noise), iterated deletion of strictly dominated strategies selects a unique equilibrium (the risk dominant equilibrium).[17]

A potentially fruitful area for future research is to use the global game approach to get sharper predictions of market structure in communication industries. This is more than an

---

the right, however, to refuse interconnection with another network, even if that network meets the criteria laid down in the guidelines.

[17] Backward induction causes equilibria that are based solely on expectations to be strictly dominated by a unique equilibrium in which agents' strategies are a function of their signals of the payoffs. The key to the argument is that the support of any agent's higher order beliefs about other agents' signals becomes arbitrarily large for a sufficiently high order of beliefs. This idea has been applied in a series of papers by Morris and Shin; see Morris and Shin (2001) for a survey.

immediate application of the Carlsson-van Damme, Morris-Shin analysis, since in those models, all agents are strategically small. An exception to this is Corsetti et al. (2000), in which there is one 'large' trader. In this model, however, the large agent's action is not observed by the small agents when they choose their action. In the network context, however, consumers do observe a network's price or output before buying. The price or quantity announcement is a public signal made by a strategic agent. Consumers use this public signal, combined with their own private signals, to form beliefs about the true state of the world (the payoff from buying the network's product, say). These beliefs are the equivalent of expectations in earlier models; but this richer setting allows expectation formation to be modelled more explicitly. Further research is needed to know whether concentrated market structures (that are supported by arbitrary expectations in earlier models) can occur in equilibrium in this case.

3.2 Pricing with negative externalities: congestion pricing and capacity expansion

By any measure, the growth in communications has been phenomenal. The major revolution has been the Internet; but other new technologies, such as mobile telephony, and ever-decreasing costs of standard technologies have also contributed. The number of hosts, the number of users, and the amount of traffic on the Internet have been doubling approximately every year since 1988. In the space of a decade, the mobile share of the world's telephone subscriptions has increased from zero to more than one in seven. Annual growth in call minutes over telephone networks lies in the range 4-15%; see Coffman and Odlyzko (1998).

The price of this success has been increasing congestion. Surfing the Web is notoriously slow during peak hours; by some estimates, 30% of Internet traffic is re-transmissions of dropped packets. (This said, it is surprisingly difficult to obtain hard evidence of Internet congestion. See Paxson (1997) for an authoritative study of the area. Many university links to the public Internet are heavily loaded, which may be why academics think congestion is a problem. It may be, however, that the general problem is not congestion, but non-responding servers; see Huitema (1997).)

This section starts by examining the economic principles for socially optimal pricing of a congestible resource. It continues by looking at congestion pricing in imperfectly competitive industries. It concludes by reviewing areas for future research.

Socially optimal pricing

Congestion is an example of a negative externality, where there is a gap between private and social valuations of use of the resource. An economist's first response, since Pigou (1920), is to implement a usage price to close this gap. There are two objectives, however, for the price to achieve. First, it should ensure the efficient level of resource usage. Secondly, it should provide the correct incentives for efficient investment in the capacity of the resource.

In principle, these two objectives are separate and may be contradictory. It is a central result of congestion pricing that under certain conditions, the two objectives are the same: the optimal congestion prices for a fixed amount of capacity automatically generate the appropriate amount of revenue to finance capacity expansion. The optimal usage price equals the total marginal cost (i.e., the sum of marginal costs) that a unit increase in usage imposes on all users. This price internalizes the congestion externality by making each user face the full costs that it imposes on all users. This price equals the marginal value of a unit of capacity; and so capacity should be expanded if and only if the revenue from congestion pricing exceeds the value of expansion.

Mackie-Mason and Varian (1995) give a nice exposition of this result. It is somewhat (but not very) technical, so here we present a numerical example of the pricing principle. N people (where N is very large) wish to use a resource e.g., the Internet to download a music file. All users have the same preferences. If the download takes D minutes and if they have to pay a charge P for using the Internet, each user's utility U is given by: $U = 100 - D - P$. They receive zero utility from not using the resource. Because of congestion, if M users are simultaneously downloading the file, the time for each download is $D = M/K$, where K is the capacity of the resource. Suppose that resource use is not priced, so that $P = 0$. In this case, the number of users N_0 who download is given by the zero utility condition $100 - N_0/K = 0$ i.e., $N_0 = 100K$. If fewer than $100K$ people download, then the utility from downloading is positive and more people will want to download. If more than $100K$ people download, then the utility from downloading is negative and fewer people will want to download. Contrast this to the socially optimal use level, N_S, which is the number of downloaders that maximizes the total utility $M(100 - M/K)$ from downloading. A straightforward calculation shows that $N_S = 50K$; the total utility from socially optimal resource use is $2{,}500K$. The negative externality of congestion therefore results in excessive use of the resource: $N_0 > N_S$, as expected.

To achieve the socially optimal use level, a social planner could charge for downloading. The price P_S that equates private and social usage is given by $100 - N_S/K - P_S = 0$, or $P_S = 50$. This price generates a revenue of $2{,}500K$ - that is, the revenue from the socially optimal congestion price equals the total welfare from using the resource in the socially optimal way.

This means that the revenue from the congestion price gives exactly the right signal for capacity expansion (investing in increasing *K*). Capacity should be increased if and only if the marginal revenue of 2,500 exceeds the marginal cost.

Market structure and oligopoly pricing with negative externalities

This establishes the economic principles behind socially optimal congestion pricing. We are also interested, however, in congestion pricing in imperfectly competitive situations. In fact, the previous discussion of positive externalities is easily adapted to this case. Recall that the central mechanism there was positive feedback: small initial differences in networks are amplified into large final differences. With congestion, exactly the reverse occurs. When two networks are of similar size, consumers will have a slight preference (all other things being equal) to join the smaller, less congested network. But then this network becomes larger; if its size goes above the other network's, then it becomes the less-preferred network. Hence any initial difference between the networks is decreased to zero by the negative feedback mechanism of congestion.

The implications of this observation can be anticipated from the earlier discussion. With pure congestion, there are no multiple equilibria; expectations do not matter; and there is no tendency for concentrated market structures. This conclusion is, however, too extreme. In fact, it is more relevant to consider the mixed case where externalities are positive when usage levels are sufficiently low, but become negative at high usage levels. (Each new web site, or the addition of information to an existing site, increases the value of the Internet to every existing user. However, as usage of the Internet grows, so does congestion.)

It would be convenient if the separate results from the models of positive and negative externalities could simply be combined to draw conclusions about the intermediate case. Lee and Mason (2001) makes a preliminary investigation of this case. They show that the consequences for market structure can be very unexpected. For example, an increase in the number of firms in the industry can increase individual firms' profits. (The reason for this is that, with a smaller number of firms, it can be that the only possible equilibria are symmetric in which profits are zero. With a larger number of firms, however, asymmetric equilibria can exist in which positive profits are earned.) This possibility arises only with both positive and negative externalities. In general, therefore, the pure positive and pure negative externality cases cannot be used to guess at outcomes in the intermediate case. Analysis of this issue is an important avenue for future research.

4. Universal Service Obligations

Regulators have long been aware of the social aspects of communication, and have been intimately involved with the various services - telecommunications, post, broadcasting etc. - since their beginnings. Due to the widespread use of these services, there are many social dimensions for regulators to cover. Initial 'public interest' arguments meant that virtually all aspects could be regulated. For example, the 1927 Radio Act in the U.S. gave federal regulators the power to issue a license to a broadcaster if they found that it was in the "public interest, convenience or necessity". The absence of any clear definition of 'public interest' means that the FCC could determine the number and identity of broadcasters, the terms and conditions of their operation, and even the content that they broadcast.

The broader social aspects of communications regulation, as well as the competition issues that we have discussing so far in this paper, can be seen in the current 'broadband debate'. Higher bandwidth services, such as high-speed Internet service, video on demand and interactive electronic commerce, have been deemed by many governments to be of fundamental importance to the development of their economy; see e.g., Oftel (1999). A particular concern is the provision of these services to residential customers (and also small businesses). This has highlighted the lack of competition in local telecommunications markets. In the U.K., broadband services are likely, in the medium-term, to be provided using enhancements (Digital Subscriber Line, DSL) to the fixed copper loop telephone network; this sector is dominated by the incumbent British Telecom, which supplies over 85% of access lines. In the U.S., local access is provided by both cable and local telephone companies; the issue there is what carriage requirements to impose on entrants to the local access market. Finally, the content that can be delivered over high bandwidth access lines has lead to regulatory initiatives such as the European Union's Action Plan on Promoting Safer Use of the Internet, adopted on January 25th, 1999.

We will not attempt to cover all aspects of regulation in this section. Instead, we will focus on the second of the two objectives stated by most telecoms regulators. For instance, the U.S. Telecommunications Act of 1996 directs the FCC to "promote competition and preserve and advance universal service". We have discussed at some length already some aspects of the promotion of competition; see section 2, for example. We will spend the rest of this section discussing the latter objective of universal service.

There are several reasons given for imposing universal service obligations (USOs). First, it is often thought that communication services (and other utilities, such as electricity, water

etc.) are necessities that should be readily available to all, simply on the grounds of equity. This argument can be supplemented with the idea that complete access to essential services stimulates economic development and growth. Thirdly, there may be significant positive externalities associated with a service (such as a communication network) that the market, left unregulated, will fail to incorporate, leading to insufficient coverage of the network. A USO may be required to correct for this market failure. Finally, although not strictly speaking a USO, it may be that social policy requires that a service be made available to a specific group of customers. For example, special conditions may apply in the provision of services to people with disabilities, in line with wider social obligations. Similarly, it may be part of educational policy to provide high-speed Internet access in schools. For these and other arguments, see Crémer et al. (1998). See also Laffont and Tirole (2000, chapter 6) and Riordan (2001) for further discussions of universal service in telecommunications.

Even if the general principles behind USOs are agreed, there is still the problem of putting them into practice. There are three aspects to this. First, what exactly should be provided and to whom? Secondly, who should be required to fulfil a USO? Thirdly, who should pay for the costs of a USO?

The exact definition of universal service is not clear. The most commonly used version refers to achieving a "minimum quality level" of a "basic package" of services to all consumers and at "affordable prices". In the case of telecommunications, this sort of statement can be found in FCC and EC communications; see the FCC's CC Docket 96-45 and the EC communication COM(96) 73. Each part of this statement is open to interpretation - what is a minimum quality level, what constitutes a basic package, what prices are affordable? Hence the FCC has listed a set of services and quality levels that are included in universal service (e.g., voice-grade access to the public switched network, touch-tone, etc.), and detailed maximum prices that can be charged for specific services, and on average across all services. This exercise is, of course, problematic. Technological progress means that the set of basic services is constantly expanding, and minimum quality levels are ambiguous (for example, wireless services allow greater mobility, but typically have lower sound quality and completion rates).

In the past, incumbent telecom operators were responsible for USOs; indeed, in the U.K., this is still the case (as mentioned in the introduction). In the U.S., USO obligations are not restricted to incumbents and universal service subsidies are paid to any company that accepts a commitment to service all consumers in its area. The subsidies are paid for typically by cross-subsidization: the income from more profitable markets (such as long-distance or

business customers) is used to cover losses incurred by charging low prices to low-income or high-cost consumers. The alternative of financing universal service subsidies through general taxation is not generally used in telecommunications, although it is used in other cases; for example, in the U.K., subsidies to winners of railroad franchises are covered from general taxes. This is despite the fact that financing from general taxation would be a cheaper (i.e., less distortionary) way to raise the required revenue, at least in developed countries.

USOs are under increasing pressure. The first source of pressure appears to be political, but actually has solid economics to back it up. A major problem with USOs is that they are blunt. A USO to cover high-cost rural areas at the same price as low-cost urban areas benefits high income rural consumers at the expense of low income urban consumers. More precisely, it may be inefficient to effect a particular objective - higher welfare for rural residents - through distorting the prices of particular services. This point has been made formally by Atkinson and Stiglitz (1976), who show that, under certain circumstances, the best way to redistribute income is through the taxation of income, not consumption.

In the Atkinson-Stiglitz model, consumers differ in their income levels (actually, in their ability levels, which affect income). Hence their result speaks most directly to the issue of subsidies to low-income consumers. It is straightforward, however, to re-interpret their model in terms of low- and high-cost consumers. One of the key conditions required for this result is that low and high income consumers have the same relative preferences for consumption goods (i.e., the marginal rate of substitution between consumption goods is independent of income). In this case, taxing consumption - effectively what occurs when the prices of telecommunications services are altered - in order to fund universal service is unnecessarily inefficient. The better way to redistribute income (which, after all, is what a universal service subsidy does) is to tax income. To encourage people to live in high-cost rural areas, the theorem suggests that a location-specific income tax break is better than offering a telecommunications subsidy.

Changes in the assumptions underlying the Atkinson-Stiglitz theorem will, of course, change the result. For example, it may be that the marginal rate of substitution between consumption goods is not independent of income. Then it may be worth taxing those goods that the rich have a relative preference for and subsidizing the goods preferred (relatively) by the poor. Nevertheless, the result is important for emphasizing that USOs must be assessed carefully for their validity and not simply accepted.

The second challenge facing USOs comes from the introduction of competition. Telecommunications markets in many countries have been opened up to competition. In the

U.S., the break-up of AT&T in 1984 allowed competition in previously monopolized markets. In the U.K., the first competitor to the previously-nationalized BT was licensed in 1982; in 1991, the market was opened further. In both cases, the idea was to use competitive forces to assist in the regulation of dominant operators. But this has consequences for the financing of USOs. USOs are supported by cross-subsidization. This cross-subsidization is sustainable while a single firm operates across the various markets. But when a second firm is able to operate, it will choose to enter the more profitable market-a process known as cream-skimming. This has three implications. First, the distortions in prices that the USO requires can lead to inefficient entry. Secondly, the subsidy required to support the USO is higher than it is when entry cannot occur; since financing the USO is distortionary, this means that the social cost of the USO is higher. Finally, USOs that come in the form of a uniform pricing requirement can have strategic effects that need to be recognised by regulators.

The first point is most clearly seen in a single market case. (The following example is taken from Armstrong (2001b).) Suppose that there is a single group of consumers with inelastic unit demand for telecommunications service. The incumbent can provide this service at cost C per consumer, giving each consumer gross utility U. The price that the incumbent charges is mandated to be P per consumer; if the consumers belong to a high-cost market, then typically $P < C$. An entrant can provide the same service at cost c, giving gross utility of u; it charges a price p, where p is not restricted (since the USO is imposed only on the incumbent). Social welfare is the sum of consumer surplus plus profit; so welfare when the incumbent serves the market is $(U - P) + (P - C) = U - C$, and when the entrant serves the market, it is $u - c$. Hence entry is socially desirable if and only if $u - c \geq U - C$ i.e., $C \geq c + U - u$. Given the incumbent's price, the entrant can attract consumers if its price satisfies $u - p \geq U - P$; that is, if $P - U + u \geq p$. Entry will occur whenever the maximum price that the entrant can charge covers its cost, that is when $P - U + u \geq c$, or $P \geq c + U - u$. Comparing this with the socially optimal condition for entry, we see that whenever P does not equal C (which is typically the case when USOs are involved), entry occurs inefficiently. When $P > c + U - u > C$, entry occurs when it is socially undesirable. When $P < c + U - u < C$, entry does not occur, even though it is socially desirable. This story can be extended to incorporate access pricing. The general moral that emerges is that when there are retail distortions due to a USO, a retail instrument should be used in combination with an appropriate access charge. Use of the access charge alone both to provide the right entry incentives and to correct the retail distortion is inferior.

When consumers are heterogeneous, with some being high-cost and others low-cost, a USO subsidy set without regard to competition will be too low. The reason is obvious: such a subsidy assumes that the operator can earn excess profits from low-cost consumers, that can be used to finance service to high-cost consumers. Competition eliminates these profits, and so increases the required subsidy. There are further effects of competition, however, studied by Choné et al. (2000) and Valletti et al. (2001). These authors show that a USO affects the way in which operators compete. In particular, a uniform pricing restriction creates linkages between markets. Depending on the nature of competition (along the lines identified in Bulow et al., 1985), this may make operators less aggressive in those markets, leading to higher equilibrium prices and deadweight loss.

The tension between universal service and competition represents a considerable challenge for regulators. A promising line of research to resolve this tension is the use of universal service auctions, in which operators bid for a level of subsidy (competition for the market), with the market structure after the auction determined by the bids in the auction (competition in the market).

5. Conclusions

In this paper, we have examined the implications of the four defining characteristics of networks: their cost structure; the strong complementarity between their components; the demand-side externalities that arise from consumption of their services; and the social obligations attached to them. Communication networks have, from their inception, been subject to close regulatory attention, due to these characteristics. Despite all of this attention, many aspects of competition between networks are still poorly understood. Add to this the rapid change in communications arising from technological progress, and you have an area that will continue to trouble regulators and interest academics for some time to come.

References

Armstrong, M. (2001a), "The Theory of Access Pricing and Interconnection," in M. Cave, S. Majumdar and I. Vogelsang (eds.), *Handbook of Telecommunications Economics*, North-Holland, Amsterdam (forthcoming).

Armstrong, M. (2001b), "Access Pricing, Bypass and Universal Service", *American Economic Review Papers and Proceedings* 91(2): 297-301.

Armstrong, M. (1998), "Network Interconnection in Telecommunications," *Economic Journal* 108: 545-564.

Atkinson, A. B. and J. Stiglitz (1976), "The Design of Tax Structure: Direct and Indirect Taxation", *Journal of Public Economics* 6: 55-75.

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

Bulow, J., J. Geanakoplos and P. D. Klemperer (1985), "Multimarket Oligopoly: Strategic Substitutes and Complements", *Journal of Political Economy* 93: 488-511.

Carlsson, H. and E. van Damme (1993), "Global Games and Equilibrium Selection", *Econometrica* 61(5): 989-1018.

Carlton, D. W. and M. Waldman (1998), "The Strategic Use of Tying to Preserve and Create Market Power in Evolving Industries", NBER WP W6831.

Cave, M. and R. A. Mason (2001), "The Economics of the Internet: Infrastructure and Regulation", *Oxford Review of Economic Policy*, Summer 2001, forthcoming.

Choi, J. P. and C. Stefanadis (2001), "Tying, Investment, and the Dynamic Leverage Theory," *RAND Journal of Economics* 32(1), 52-71.

Choné, P., L. Flochel and A. Perrot (2000), "Universal Service Obligations and Competition", *Information Economics and Policy* 12: 249-259.

Coffman, K. G. and A. Odlyzko (1998), "The Size and Growth Rate of the Internet", *First Monday* 3(10).

Corsetti, G., A. Dasgupta, S. Morris and H. S. Shin (2000), "Does One Soros Make a Difference? The Role of a Large Trader in Currency Crises", *Mimeo*, Department of Economics, Yale University.

Crémer, J., P. Rey and J. Tirole, (2000) "Connectivity in the Commercial Internet", *Journal of Industrial Economics*: 48(4): 433-472.

DeGraba, P. (2000), "Bill and Keep at the Central Office as the Efficient Interconnection Regime," OPP Working Paper 33, Federal Communications Commission, Washington, DC.

Dessein, W. (2000), "Network Competition in Nonlinear Pricing," *mimeo*, GSB Chicago.

Diamond, D.W. and P. H. Dybvig, (1983) "Bank Runs, Deposit Insurance, and Liquidity", *Journal of Political Economy* 91(3): 401-19.

Farrell, J. and M. Katz, 2000, "Innovation, Rent Extraction, and Integration in Systems Markets," *Journal of Industrial Economics* 48(4): 413-432.

Farrell, J. and G. Saloner, (1985) "Standardization, Compatibility, and Innovation," *RAND Journal of Economics* 16: 70-83.

Farrell, J. and G. Saloner, (1986) "Installed Base and Compatibility: Innovation, Product Preannouncements, and Predation," *American Economic Review* 76: 940-955.

Farrell, J. and G. Saloner, (1992) "Converters, Compatibility, and the Control of Interfaces," *Journal of Industrial Economics* 40(1): 9-35.

Gans, J. (2001), "Regulating Private Infrastructure Investment: Optimal Pricing to Essential Facilities," *Journal of Regulatory Economics* 20(2): 167-189.

Gruber, H. and T. M. Valletti (2001), "Mobile Telecommunications and Regulatory Frameworks," in G. Madden and S. Savage (eds.), *The International Handbook of Telecommunications Economics*, Edward Elgar, Aldershot (forthcoming).

Katz, M. L. and C. Shapiro (1985), "Network Externalities, Competition, and Compatibility," *American Economic Review* 75(3): 424-40.

Klemperer, P. (1995), "Competition when Consumers have Switching Costs: An Overview with Applications to Industrial Organization, Macroeconomics, and International Trade," *Review of Economic Studies* 62(4): 515-39.

Huitema, C. (1997), *The Required Steps towards High Quality Internet Services*, Unpublished Bellcore Report.

Jeon, D. S., J.-J. Laffont and J. Tirole (2001), "On the Receiver Pays Principle," *Mimeo*, IDEI, Toulouse University.

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

Laffont, J.-J., P. Rey and J. Tirole (1998), "Network Competition: I. Overview and Nondiscriminatory Pricing; II. Discriminatory Pricing," *RAND Journal of Economics* 29: 1-56.

Lee, I. H. and R. A. Mason (2001), "Market Structure in Congestible Markets", *European Economic Review* 45(4-6): 809-818.

MacKie-Mason, J. K. and H. Varian (1995), "Pricing Congestible Resources," *IEEE Journal of Selected Areas in Communications* 13(7): 1141-49.

Mason, R. A. (1999), "Compatibility between Differentiated Networks," University of Southampton Discussion Paper in Economics and Econometrics, No. 9909.

Matsuyama, K. (1991), "Increasing Returns, Industrialization, and Indeterminacy of Equilibrium," *Quarterly Journal of Economics* 106(2): 617-50.

Matutes, C. and P. Regibeau (1988), "'Mix and Match': Product Compatibility without Network Externalities," *RAND Journal of Economics* 19(2): 221-34.

MMC (1998), *Cellnet and Vodafone*. Monopolies and Merger Commission, London.

Morris, S. and H. S. Shin (2000), "Global Games: Theory and Applications", *Mimeo*, Department of Economics, Yale University.

Nalebuff, B. (2000), "Competing Against Bundling," WP 7, Yale School of Management

Odlyzko, A. (2000), "The History of Communications and its Implications for the Internet", *Mimeo*, AT&T Labs.

OECD (2000), *Cellular Mobile Pricing Structures and Trends*. DSTI/ICCP/TISP(99)11/final, Organisation for Economic Co-operation and Development, Paris.

Oftel (1999), *Access to Bandwidth: Proposals for Action*, Consultation document issued by the Director General of Telecommunications, London.

Paxson, V. (1997), Measurements and Dynamics of End-to-end Internet Dynamics, Ph.D. thesis, Computer Science Division.

Pigou, A. C. (1920), *The Economics of Welfare*. London: Macmillan.

Riordan, M. (2001), "Universal Residential Telephone Service", in Cave, M., S. Majumdar and I. Vogelsang (eds.), *Handbook of Telecommunications Economics*, North-Holland, Amsterdam (forthcoming).

Shapiro, C. and H. Varian (1999), Information Rules, Harvard Business Studies.

Valletti, T. M. (1999), "The Practice of Access Pricing: Telecommunications in the UK," *Utilities Policy* 8(2): 83-98.

Valletti, T. M. (2000), "Switching Costs in Vertically Related Markets," *Review of Industrial Organization* 17(4): 305-409.

Valletti, T. M., S. Hoernig and P. P. Barros (2001), "Universal Service and Entry: The Role of Uniform Pricing and Coverage Constraints," *Journal of Regulatory Economics* (forthcoming).

Whinston, M. D. (1990), "Tying, Foreclosure, and Exclusion," *American Economic Review* 80: 837-859.

Willig, R. D. (1979), "The Theory of Network Access Pricing," in H.M. Trebing (ed.), *Issues in Public Utility Regulation* Michigan State University Public Utility Papers.