Competition in services or infrastructure-based competition?*

By Mats A. Bergman
The Swedish Competition Authority, 103 85 Sthlm
and Stockholm University

September 16, 2004

1. Introduction

A perfectly competitive market has uniquely favourable characteristics: firms produce efficiently and sell to consumers at marginal costs. On the other hand, no real market has all the properties that are necessary for perfect competition, such as perfect information and price taking by every firm and every consumer.

At best, some markets can be approximately perfectly competitive. In practice, most markets deviate quite substantially from the perfectly competitive ideal. A possible policy response is to introduce regulation, in order to correct for the failures of the market. However, regulation gives rise to inefficiencies. No matter what regulatory model is chosen, there will be disadvantages. Depending on how well an unregulated market would function and on how efficient regulation would be, the best course of action may either be to accept a certain degree of inefficiency in an unregulated market – or to introduce regulation.[1]

There are many ways to make regulatory mistakes, some more costly than others.[2] The regulator's situation is asymmetric in an unrewarding way: it is easy to make mistakes that will be immensely costly, while it is difficult to make improvements that will have even relatively modest payoffs.

The telecom market is, and has for a long time been, quite extensively regulated, although the *nature* of the regulation has changed dramatically. Government ownership of national telecom

* All opinions expressed in this article are my own and do not necessarily reflect those of the Swedish Competition Authority. I am grateful to Lars Hultkrantz and Mikael Ingemarsson for insightful comments. This report was commissioned by the Swedish National Post and Telecom Agency.

[1] This is of course a simplification. Regulation can be more or less intrusive and most economists would argue that some regulation is necessary in all markets. General commercial law, competition law and consumer-protection legislation, for example, can be seen as different forms of regulation that enhances the efficiency of the market. In the following, I will use the term "regulation" to denote sector-specific "economic" regulation that limits the firms' freedom to set prices and quantities and make decisions to enter or exit markets, et cetera.

[2] E.g., California's electricity crises and rail regulation in Britain; see Bergman, 2002.

monopolies was the predominant regime in Western Europe before the 1990s, while in the US consumer prices were regulated. Recently, the trend in most industrialized countries, including Sweden, has been towards *access* price regulation. In some respect, this so-called deregulation has substantially *increased* the apparent quantity of regulation. At the same time, the telecom firms' latitude to make business decisions has also increased.

As competition has increased in the telecom markets, one can envision a future where access regulation and other types of sector-specific regulation can be dismantled. That is, the market for telecommunications may eventually become an "ordinary" market, where general competition rules (and other general legislation) will be enough to maintain competition. If this were to happen, we would be able to benefit from effective competition, without the costs and other disadvantages of regulation. It appears that such a course of event would require effective competition in the provision of infrastructure, not just competition in services, since one firm (the owner of the infrastructure) would otherwise inevitably be in a monopoly position. In several areas of the telecommunication market, competition in infrastructure has indeed developed. Due to technological progress, long-distance connection, that used to be natural monopoly, is now competitively provided. The evolution of mobile telephony has resulted in overlapping networks and there is even some competition for fixed access, in particular in business districts, but also more generally for broadband access.

A possible conclusion, then, is that infrastructure-based competition is both feasible and preferable to a regime with access regulation and competition in services only. An additional argument in support of this position is that regulation may distort the firms' investment incentives – if regulation is too strict (if access fees are too low), there will be too little investment.

On the other hand, competition in infrastructure requires duplication of assets, which may be inefficient if one or two sets of assets has enough capacity to serve the whole market. This suggests that policy makers will face a trade-off between competition and returns to scale. In addition, because of the special nature of the telecom market, sector-specific regulation may be necessary even if there are multiple infrastructures. For example, under the 2003 EU Electronic Communication directive, telecom operators are required to provide termination access at cost-based prices, no matter the number of competing networks. This means that the choice may not be one between access regulation and unregulated competition, but between *one-way access* regulation and *two-way access* regulation.

One-way access regulation is the traditional regulation of a bottleneck owned by one firm, such as the local loop of the fixed-line telecom network, owned by the incumbent operator (the former national monopoly). The downstream competitors need access to facilities controlled by the incumbent firm, but the incumbent has no need for access to the competitors' facilities. Two-way access regulation requires two or more firms to provide access to each other. For example, competing network-owning mobile operators need access to each others' nets, so that calls originated in one network can be terminated in another. Outside the telecom sector, two-way access regulation is less common, except in the financial markets. Just as telecom operators need access to each others' customers, banks need to access each others' account systems, in order to process debit or credit transactions. An important difference, however, is that while telephone calls require instantaneous access, bank transaction can often be executed with a delay. This makes it possible to use multilateral systems, or clearing and settlement institutions, such as giro centrals and systems for processing card transactions.

It is useful to make a distinction between markets where the access provider competes for customers with the firm(s) that seek access, and those where there is no such competition. An example of the former is a market with a vertically integrated incumbent that competes in the telecom services market with one or more entrants. An example of the latter is a market with a vertically separated telecom network operator, which does not compete in the downstream services market. Table 1 provides a classification scheme that distinguishes both between one-way access and two-way access, and between markets with and without competition for customers (in the above sense).

Table X.1. Four possible access regimes in telecom markets

	Competition for subscribers	
	Yes	No
One-way access	Origination and termination of fixed lines; LLUB	Vertically separated infrastructural companies (e.g., Banverket and Stokab)
Two-way access	Mobile-to-mobile interconnection	Intl. interconnection; Fixed-to-mobile interconnection

Of course, there are many aspects of the regulatory regime that are not captured by the categories of the table. Two-way access regulation can, for example, be symmetric or asymmetric. The latter may be relevant in markets where two or more mobile operators own networks, but where one of them has a dominant position. The nature of the optimal regulatory scheme may also depend on whether consumer prices are linear or non-linear, on whether there are receiver payments or not and on the regime for setting retail prices (market-determined prices or regulated prices). Armstrong (2002) argues strongly that policy makers should make greater use of output taxes, levied on incumbents and entrants alike, in order to increase efficiency. The main difference between an output tax and an access fee is that the former has to be paid irrespective of whether the operator uses its own infrastructure or infrastructure owned by someone else, while access fees will only have to be paid in the latter case. The main advantage of introducing output taxes is that they give the regulator one more "instrument" to achieve efficiency, in particular when the regulator (or the legislator) simultaneously pursues other objectives than pure efficiency, such as universal service. Output taxes will then, in fact, be fees to fund universal-service obligations.

The choice between competition in services and infrastructure-based competition is a complex one. At the bottom lies the choice between the benefits of free competition and the benefits of returns to scale. Infrastructure-based competition offers the potential of less regulation and hence less regulation-induced inefficiency, such as distorted investment incentives, lobbying and pure bureaucracy costs. Service-based competition, on the other hand, allows the industry to realize greater returns to scale. Furthermore, introducing competition in infrastructure may not lead to deregulation, but only to re-regulation: from one-way access regulation to two-way access regulation. It follows that the reduction in regulation-induced inefficiencies from introducing competition in infrastructure may turn out to be smaller than many think. On the

other hand, technological developments may have the effect that returns to scale in infrastructure are not as big as they appear to be. Alternatively, competition *for* the infrastructural market may be feasible, even if competition *in* the market is not possible.

In the end, we are faced with a choice between two slightly different regulatory regimes, one of which (service competition) allows us to economise on investments in infrastructure and one of which (infrastructure-based competition) is a little less interventionistic. What the best choice in a particular situation is cannot be deduced in the abstract. That depends on the available technologies – i.e., on whether there are significant returns to scale or not – and on the relative merits of the alternative regulatory schemes. With one-way access, the main concern will be that the incumbent will be able to foreclose smaller rivals, while under two-way access (multilateral) access agreements can be used to achieve coordination of retail (consumer) prices. The regulator's ability to combat these problems will, in turn, depend on a number of factors, such as the operators' market shares and the structure of retail prices.

One conclusion of this paper is that the vision of a "sunset" for telecom regulation may be at least partially misleading. According to this vision, to which EU Commission has alluded in the process of launching the E-com directive, sector-specific telecom regulation will eventually become unnecessary: when facilities-based competition has evolved, general competition rules will be sufficient. However, in the presence of substantial returns to scale on the supply side (the cost of building networks), as well as on the demand side (network effects), effective competition may never evolve. Hence, the regulator (or the legislator) must make up its mind as to where and how facilities-based competition should be encouraged, while maintaining a regulatory regime, at least for two-way access and possibly also for one-way access.

Another conclusion, by no ways novel, is that regulation must be designed so as not to distort investment incentives. In particular, this suggests that while strict access regimes to old monopoly networks may be warranted, one must be careful not to impose too strict access regimes on networks that have been built in competition. However, a distinction can be made between access that is motivated by returns to scale on the supply side and on the demand side, respectively. An example of the former is access for origination, which can sometimes be motivated if the cost of duplicating the network is large. An example of the latter is access for termination, which is also motivated by network effects. Because of network effects, relatively strict (two-way) access regimes for termination can perhaps be justified, even in the absence of a dominant incumbent. However, in order not to distort investment incentives, one must be much more careful with imposing (two-way) access regimes for origination in competitive markets. The former part of this insight is reflected in the E-com directive, where each network is presumed to be a separate relevant market for termination. On the other hand, there is a potential risk that the latter part of the insight is ignored – i.e., that the directive is used too aggressively in favour of operators seeking access for origination.

The next section discusses returns to scale (or scale economies) in some depth. Section 3 focuses on the so-called bottleneck problem, which arises when there are two or more interdependent stages of production, with different degrees of scale economies. Section 4 discusses short-run and long-run competition, investment incentives, and some aspects of regulatory failure. Section 5 deals with the Efficient Component Pricing rule and efficient and inefficient bypass (duplication of infrastructure). Section 6 makes comparisons with the electricity and postal markets, and section 7 concludes.

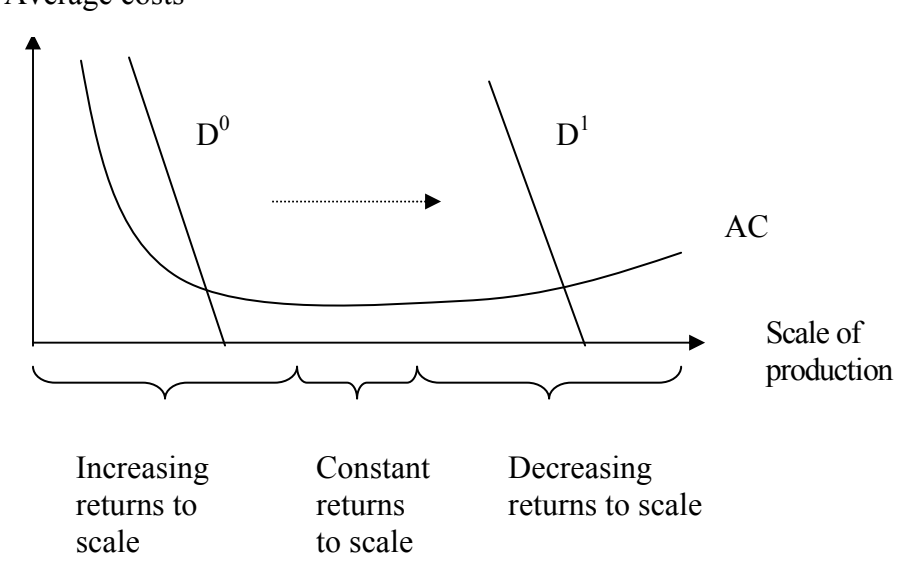
2. Natural monopolies, returns to scale and network effects

When there are increasing returns to scale, one sometimes has to compromise between two means for achieving efficiency: large-scale production and competition. This choice is often relevant in telecom markets, as this is an industry characterised both by significant returns to scale and strong network effects, and as, at the same time, it is apparent that the introduction of competition into this market has brought a number of benefits.

2.1 Returns to scale and the definition of natural monopolies

The concept of natural monopoly is linked to the concept of returns to scale, although the two concepts are not synonymous. If the average cost of some production process is falling as the scale increases, then the process is said to be characterized by increasing returns to scale (or economies of scale). The situation is illustrated in Figure 1. Note that returns to scale typically depends on the scale of production. I.e., a process that is characterized by increasing returns to scale when production is relatively small may be characterized by *diseconomies* of scale (or decreasing returns to scale) at large production scales. In the figure, this is illustrated by an average cost curve that first falls, then is constant for an intermediate range and then begins to rise.

Figure 1. Returns to scale

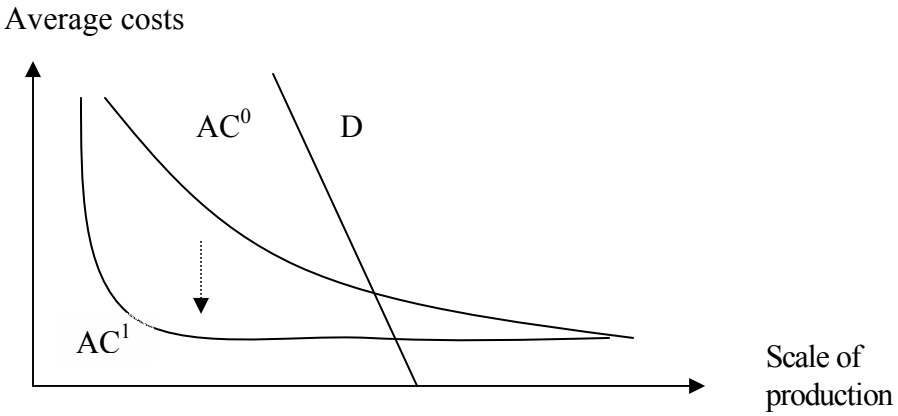


Note also that as long as the average cost is falling, marginal costs must be below average costs, while the opposite holds when the average cost is rising.

An industry may be a natural monopoly when production is small, while *not* being a natural monopoly when production is large. So, for example, a gas station may be a natural monopoly in a small town, while the gasoline retail market is not a natural monopoly in a big town. In a given mobile telephony market, mobile telephony infrastructure may initially – while the number of customers is still small – be a natural monopoly. As the number of customers increases, the market may eventually develop so that it no longer is a natural monopoly. In the figure above, this can be illustrated with a demand curve that shifts to the left.

An industry's transformation from a natural monopoly into an industry that is not a natural monopoly may also be caused by a developing technology. This is illustrated in Figure 2. With the initial technology, average costs falls with scale, as illustrated by the AC^o-curve. However, a new technology is introduced, resulting in average costs AC^1. Initially, the industry is a natural monopoly when demand is given by D. After the technology has changed, this is no longer so.

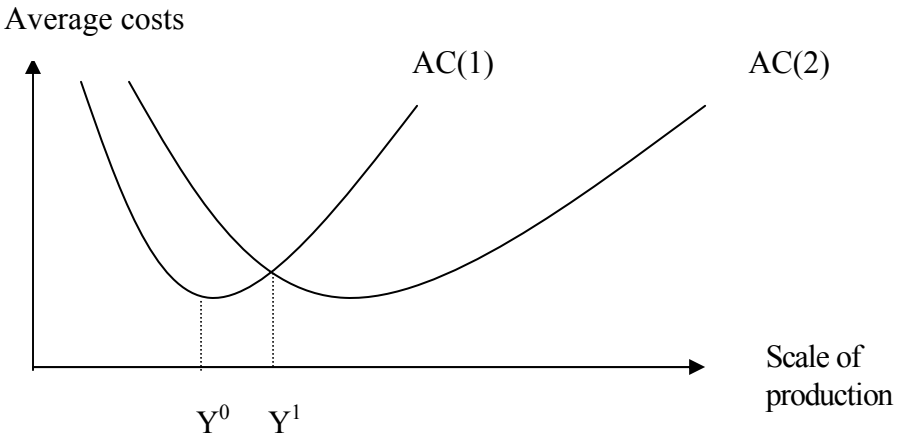
Figure 2. Returns to scale and technological change



The causes of economies or diseconomies of scale will be discussed below, but let's return to the relation between economies of scale and natural monopolies. We have seen that returns to scale may vary with the scale of production and that whether an industry is a natural monopoly or not may depend on the scale of production. It may be conjectured that an industry is a natural monopoly for those levels of production where there are positive economies of scale, while it is not a natural monopoly for those levels of production where there are diseconomies of scale. Although this may be approximately correct, matters are slightly more complicated.

For some levels of output, it may be the case that total costs are lower if production is concentrated to on firm, even if average costs are rising due to diseconomies of scale. Then, for this these output levels, the industry is a natural monopoly while simultaneously the industry exhibits diseconomies of scale. The situation is illustrated in Figure 3.

Figure 3. Returns to scale and natural monopolies



The curve AC(1) illustrates the average cost (as a function of production) if there is one firm and the curve AC(2) illustrates the average cost if production is divided equally between two firms. For production levels below Y^0, there are positive returns to scale, while for larger production levels there are diseconomies of scale (average costs will rise as production

increases). Despite this, for quantities slightly larger than Y^0, costs will be higher if production is divided between two firms rather than concentrated to one firm. Only for quantities larger than Y^1 will it be efficient to divide production between two firms.

Natural monopolies are defined along these lines. In economists' jargon, an industry is a natural monopoly if, over the relevant range of production, the costs of production are *sub-additive*. In plain English: if it is less costly to have one firm producing, rather than dividing production between two or more firms, then the industry is a natural monopoly. Hence, in Figure 2, the industry is a natural monopoly below Y^1, but not above that level. In contrast, there are positive returns to scale below Y^0 and decreasing return to scale above that level. Note also that even if the cost function (i.e., the technology) does not change, an industry may eventually, as the consequence of growing demand, stop being a natural monopoly. Such a situation is shown in Figure 2 above.

2.2 Sources of returns to scale

Economies of scale can arise both on the supply (or cost) side and on the demand side. Supply-side economies of scale are related to the production technology as such, while demand-side economies of scale are more relate to the characteristics of the product and to the consumers' and producers' desire to interact – e.g., making phone calls to each other.

A number of reasons suggest that, in general, there should be increasing returns to scale in production:

- There may be fixed firm-specific costs (e.g., management, R&D and firm infrastructure) that do not rise as the scale of production increases.
- Increased scale may allow the firm to shift towards more efficient technologies - typically more automated technologies, with relatively higher fixed costs and lower variable costs.[3]
- A higher level of production will allow employees to become more specialised and will allow individuals and firms to move down the learning curve.
- So-called economies of massed reserves will allow firms to economise on production equipment, as random breakdowns or idiosyncratic demand and supply fluctuations will have less impact.

However, these sources of scale economies will eventually peter out, and diseconomies will set in, such as increasing managerial costs due to the complexity of the operation, agency problems[4] and, in many industries, transportation costs.[5]

In telecom, specifically, the main source of scale economies appears to be the infrastructure. There are large fixed costs associated with building an infrastructure for fixed telephony. Clearly, these costs will be higher if the network is duplicated, but the per-subscriber cost is also likely to be high in areas with few subscribers. The infrastructure is of fundamental

[3] Cf. the so-called two-thirds rule, shown to apply for many chemical and metallurgical processes.
[4] That is, incentive problems due to asymmetric information between the firm's owners and its management, and between different levels in the internal hierarchy. In other words: in a big organization, your ultimate boss will not know exactly what you are doing and why, so you may as well use some of your time and effort to further your personal interests.
[5] See Tirole, 1988, section 1.2 and Scherer and Ross, 1990, chapter 4.

importance also for the cost structure of mobile telephony operators. In sparsely populated areas, the cost will be driven by the need to get geographical coverage, while in densely populated areas (or areas with a lot of traffic), base stations will have to be built much closer to each others, in order for them to be able to handle enough traffic. As will be discussed in the next sub-section, these particular types of scale economies can also be seen as "returns to density".

Sung and Gort (2000) empirically estimate the returns to scale on a sample of Local Exchange Carriers (LECs) in the US. They find only small evidence of returns to scale – and that the two largest LECs actually operate on a scale where diseconomies of scale has set in. Fuss and Waverman (2002) report contradictory empirical findings on returns to scale: some studies find positive returns to scale and some find negative returns. According to Falch (2001), there is no consensus on the level of scale economies in telecom. Falch makes the point that technical estimates of returns to scale may give biased results, as such estimates will be conditional on the technology chosen. In particular, large operators are likely to choose technologies with large returns to scale, even though other technologies may be equally efficient at the firm's current scale. Falch presents some simple comparative measures of productivity for a sample of operators from countries of different sizes and finds no clear evidence of positive returns to scale. That is, telecom operators from large countries (i.e., operators with a large number of customers) are no more efficient than operators from small countries (operators with relatively few customers).

2.3 Returns to density

Since the product that the telecom industry provides is a communication service between different geographical locations, it is possible to make a distinction between returns to scale and returns to density (or economies of density). There are returns to density when, on a given route or line or within a given geographical area, average cost falls as traffic on that route or line increases or as transactions volumes in that area grows. There are returns to scale when average costs falls as the number of routes or lines served by one company increase, or as the firm expands into a larger geographical area. (It is possible to view economies of density as a special case of economies of scale: the former type of economies appear when the scale of production increases *within* a given geographical area, but not when the scale of production increases by expanding that area.)

To be concrete, if there are returns to density in mobile telephony, then it is efficient to have one operator in a given area, but it is not necessarily efficient that the area covered by each operator is large. Without loss of efficiency, each mobile operator may be the monopoly provider in relatively small areas, such as cities, but there may be several operators in the domestic market. Similarly, if there are returns to density but not (other) returns to scale, mobile operators in small countries, with few customers, may be equally efficient as mobile operators in large countries, with many customers, at least if the small country is as densely populated as the large countries. Conversely, if there are returns to scale in mobile telephony, but not returns to density, then it is efficient to have large companies, but without loss of efficiency, the companies can all be active in all geographical areas.

Of course, there may be both returns to scale and returns to density. At least in sparsely populated rural areas, it may be efficient that just one company builds infrastructure for mobile telephony – implying returns to density. At the same time, it may be efficient that one

company has a large number of customers in all parts of a large country – or even in different countries.

There are reasons to believe that returns to density are different for fixed and mobile telephony and different between areas with dense and sparse traffic. For example, returns to density in mobile telephony may be exhausted in central business districts (since the number of base stations may have to increase roughly in proportion to traffic volumes), while being substantial in rural areas. How large returns to density and returns to scale are in practice is an empirical question. Unfortunately, there appears to be a paucity of empirical studies that clearly makes a distinction between returns to scale and returns to density in telecom.

Viscusi *et al.* (2000)[6] describe how the cable TV market is characterised by significant economies of density, while the economies of scale are quite small. If a cable system's market penetration increases from 40 to 80 per cent, then average costs decline by over 40 per cent; if the number of subscribers doubles from expanding the geographical area covered, then costs fall by approximately five per cent.[7] The cable TV industry appears to have some characteristics in common with fixed telephony, suggesting that these numbers may be of some relevance for the telecom market.

If there are significant returns to density, then competition *in* the market should be avoided. This, in turn, suggests that either the industry should be regulated, or else there should be competition *for* the market – for example through franchise bidding. (See section 3.2) However, if at the same time returns to scale are relatively modest, then horizontal separation can be option, although this would perhaps require vertical separation too.[8]

2.4 Network benefits

In addition to returns to scale from the supply (or cost) side, there may be important returns to scale from the demand side; such effects are referred to as network effects,. In a market without network effects, the consumer cares only about his or her own level of consumption (and, of course, for the price, the quality of the product et cetera). In a market with network effects, the consumer cares - directly or indirectly - also for other consumers' levels of consumption. The benefit each person derives from his or her consumption increases as the number of other consumers increases, i.e., with the size of the market. There is a positive scale effect that comes from increased per-consumer benefit, not from a reduction of costs. The network effect can in fact be seen as a positive externality between consumers.

In the simplest setting, the number of other consumers of the same product has a direct effect on the (marginal) utility of consuming a unit of the product. An example would be telephones or faxes: a given consumer's utility from having a phone or a fax increases with the number of other consumers that also have phones and faxes, respectively. This type of network effect is sometimes called one-sided network effect.

---

[6] See pp. 412-416, which are based on original research by Webb (1983), Noam (1985), and Owen and Greenhalgh (1986).

[7] The latter figure extrapolates and interprets the original statement: that average costs fall by 0.5 % as the number of subscribers increase by 10 % in a cross- section study of approximately 4000 cable systems.

[8] See Section 3.2. An interesting comparison is ATM (Automated Teller Machine) networks. In the US, the ATMs are sometimes owned by small independent firms which, in a sense, are specialised in owning and operating infrastructure. These firms contract with banks, while the ultimate customers, i.e., people making cash withdrawals, typically are unaware of who owns the ATMs they use.

A somewhat more complex situation arises when there are two types of agents that interact on one "platform". Either type cares for the number of agents of the other type that uses the platform, but not (directly) about the number of agents of its own type that does so. Some examples are buyers and sellers in advertising markets and marketplaces for trading (e.g., stock markets), as well as matchmaking markets (such as dating agencies, real estate agents and business-to-business websites). A buyer does not directly benefit from the presence of other buyers - and may indeed suffer from the increased competition for the sellers' product that additional buyers bring. On the other hand, the buyer derives benefit from the presence of additional sellers, while the sellers derive benefit from the presence of additional buyers. Hence, buyers may *indirectly* benefit from there being a large number of other buyers, as this will attract a large number of sellers - and vice versa. This phenomenon is known as two-sided network effect. Another example is the market for operative systems for personal computers: the operative system is a platform that is used by software manufacturers and by users of personal computers. An operative system such as Windows, that has a large installed base of users, is an attractive platform for software developers. Conversely, if a large number of applications have been developed for an operative system, that system will be attractive for new users. More generally, many manufacturing standards (computers and hardware, CD players and CDs, et cetera) and communication protocols are examples of markets with two-sided network effects. Yet another example is shopping malls, which must attract customers as well as retailers.

Sometimes a third category of network effects is identified: indirect network effects in one-sided markets. Possible examples are public-transport networks and (single-bank) Automated Teller Machine (ATM) networks. A higher number of passengers and a higher number of cardholders on the ATM network, respectively, will result in more frequent departures and a denser (or wider) ATM network. This increases welfare for the average customer, even though congestion effects may imply that the direct effect on a given passenger's utility of another passenger may be negative, and similarly for an additional ATM cardholder. This type of network effect is very reminiscent of ordinary scale (or density) economies: as the number of customers in a retail outlet increases, the retailer can expand its range of products, it can extend opening hours and it can often reduce prices. Similarly, the manufacturer of some widget will often be able to reduce average costs when the scale of production increases. However, if different banks join the same ATM network, or if different airlines, say, use the same airport, then this can be seen as an example of a market with a platform and two-sided network effects.[9]

In network markets, competition *between* networks must be distinguished from competition *within* systems. Examples of *inter*-network competition (competition between firms using different networks) are PC computers vs. Apple computers and American Express credit cards vs. Visa vs. Mastercard. Examples of *intra*-network competition (competition between firms that use the same network) are competition between various PC producers; competition between banks offering to process Visa transactions for merchants; and competition between different telephone operators. (*Note also that the distinction between intra-network competition and inter-network competition is not absolute, as there is often a degree of compatability even between supposedly non-compatible systems.)*

[9] Armstrong (2004) provides further examples and analyses network effects in two-sided markets in a general setting.

Guibourg (2001) presents results that suggest that network effects are more important than cost-side returns to scale in the market for payment cards.[10] It is likely that this is true also for telephone markets. In fact, the necessity of interconnection is taken for granted in the telecom industry, both in fixed and mobile telephony. Fixed telephony operators have to provide interconnection both for origination and termination, while mobile telephony operators have to provide interconnection for termination.[11]

In other industries, interconnection is not always taken for granted. Manufacturing industries, for example, often prefer to use proprietary standards if they see a chance to dominate an industry. Although industry-wide standardization and "open source" arrangements are common in practice, this is perhaps due to the absence of a clear dominant.

2.5 The benefit of competition

Just as fundamental as economies of scale, are the benefits of competition. When competition is lacking, one or a few firms will possess market power, which in turn has four main adverse consequences.

- Welfare is transferred from consumers to producers.[12]
- As the price rises above the competitive level, demand will fall below the optimal level – i.e., there will be allocative inefficiencies.
- A low competitive pressure is generally believed to result in sub-optimal effort levels and X-inefficiencies. (I.e., weak cost control will result in too high costs.)
- The existence of a monopoly profit may trigger socially costly lobbying for the favoured position, as well as other types of rent-seeking behaviour.

Although regulation can mitigate problems of the first and the second type, there is a substantial risk that it will not properly address problems of the third and fourth type. In addition, regulation brings new problems, such as regulatory risks (the risk that investment incentives *et cetera* will be reduced, because the regulator may be tempted to exploit the regulated firm after it has sunk the investment cost) and the direct costs of regulation.[13]

Firms have a strategic interest to overstate economies of scale and to downplay the benefits of competition. This is so, because a reduction in the number of competitors is typically beneficial for the industry and negative for consumers, while economies of scale will tend to benefit both categories. Hence, it is tempting to appeal to economies of scale also in situations where the true rationale is a desire to reduce competition.

Some observers argue that the introduction of competition in previously regulated markets normally gives rise to cost savings and price reductions in the 25-75 % range (Winston, 1998, and a number of OECD studies, referred to in Gonenc and Nicoletti, 2000). Based on an extensive review of the empirical literature on deregulation, Bergman (2002) arrives at the

[10] In a cross-country study, Guibourg finds that the per-capita number of card transaction rises quickly following reforms that make previously incompatible card systems compatible, while she also finds that given the number of mutually incompatible systems in a country, the absolute size of the system has no effect on the per-capita usage – suggesting that size has no strong effect on costs.
[11] Prior to the "deregulation" of the Swedish telecom market, entry into the market was in principle allowed, but in practice impossible, since the incumbent, Televerket, could chose not to interconnect. (Bergman, 2002.)
[12] Although, strictly speaking, this will only reduce welfare if we value consumer surplus higher than producer profit, welfare transfers from consumers to producers are normally considered to be negative.
[13] See Bergman, 2002.

conclusion that a more realistic prospect is savings in the 5-10 % range.[14] However, it should be noted that these studies focus on the effect of "deregulating" previously regulated industries. Given that the pre-existing regulation was not completely counter-productive, the difference between monopoly markets and competitive markets can be expected to be larger – perhaps much larger.

Another line of research focus on the difference between (unregulated) monopoly markets and markets with a larger number of firms. The general conclusion is that prices fall as the number of competitors increases, with the greatest price difference between monopoly and duopoly and then successively smaller changes. How much prices falls varies from industry to industry. Note, also, that these studies do not focus on industries that are *legal* monopolies. Even if a firm has a de facto monopoly position, the threat of potential entry may often be an effective competitive constraint. The extent to which a monopoly is able to increase prices above the competitive level can be expected to depend on the level of entry barriers (as well as on the price elasticity of demand and so on). If entry barriers are high, a monopoly can increase prices substantially, while if entry barriers are low, potential competition will be effective.

If a firm were given a legal monopoly status – or if entry barriers were very high – and in the absence of price regulation, the effect on prices would probably be quite substantial. A possible way to estimate the likely price effect of such market configurations is to look at the effect of cartels. Carlton and Perloff (2004), based on Posner (2003), report that a sample of international cartels resulted in price increases of 30-100 per cent. Connor (2003), based on a sample of 70 cartels, reports an average overcharge of 28 per cent. However, in neither of these studies were the researchers able to actually *measure* the overcharges. Instead, they relied on various methods to estimate the cartels' actual effect.

2.6 The trade-off between returns to scale and competition

It may sometimes be justified to introduce competition even if an industry is a natural monopoly. Whether an industry – or a particular stage in the production chain – should be a monopoly or not depends on the benefits of competition *relative to* the magnitude of the economies of scale. It may simply be the case that the benefits of competition are big enough to make some duplication worthwhile. However, when an industry is a natural monopoly, the market may need a helping hand, in the form of regulation, for competition to be established at all.

Arguably, there are significant returns to scale in the telecom industry, from the cost side as well as from the demand side (i.e., network effects). Possibly, the latter are the most important, but the regulatory requirement of interconnection between telephone operators suggests that the bulk of such network effects *will* be exploited. In a sense, the industry becomes less of a natural monopoly because of the interconnection regulation. In the absence of interconnection, it may be efficient that all consumers patronize the same producer; if there

[14] The large effect of deregulation found in some studies appears to stem from two types of shortcomings in the empirical research design. One is that a falling trend in costs and prices that existed even before the deregulation is not accounted for (e.g., in telecom and rail freight). The other is that short observation periods are used to estimate changes in the rate of productivity growth – and that these estimates are used to extrapolate deregulation gains far beyond the period of observation.

is full interconnection, there will no longer be any *demand-side* reasons for a monopoly. Note, however, that regulation of termination charges may be necessary to achieve this situation.

However, in telecom markets there appears to be a non-trivial trade-off between, on the one hand, *supply-side* returns to scale and density and, on the other hand, competition. Depending on the particular circumstances at hand, the policy maker can try to achieve efficiency through a number of different mechanisms, such as regulation of consumer prices and access prices, government ownership, monopoly franchises and vertical separation. The relative merits of these policies depend, i.a., on the magnitude of the returns to scale, on how amenable the industry is to efficient regulation, on the elasticity of demand and on what benefits can be obtained from introducing competition. An important insight, however, is that different stages of production may have different degrees of scale economies. In particular, returns to scale (or density) may be very large in providing infrastructure, while the returns to scale in service provision may be much smaller, or even negative. The infrastructural stage may then serve as a *bottleneck* that limits the degree of competition for the whole industry. Absent this bottleneck, competition would perhaps be vital even without regulation. From a policy perspective, regulation may have to accommodate such differences between the industry's successive production stages. The bottleneck problem, as well as possible policy responses, is the issue of the next section.

3. Multi-stage production and the bottleneck problem

In almost all industries, the production process consists of several distinct stages. A farmer, for example, buys seed**,** fertilizers and farm equipment in order to produce grain. The inputs used by the farmer are produced by specialized firms, which in turn buy energy, basic commodities and equipment from other firms. The farmer's grain may be sold to a milling company, which sells flour to firms that makes pasta or bread, which in turn sell to food wholesalers and retailers. Similarly, the production of a car involves, among others, the activities of producers of basic commodities, possibly several stages of auto-part manufacturers that produce successively more complex parts of the car, auto manufacturers (who's activities involves, i.a., design, assembly, marketing and distribution) and auto retailers.[15]

In most industries, different firms specialize in different stages. Farmers rarely produce artificial fertilizers or mill their own grains, while auto manufacturers to an increasing extent buys complex parts from external sources, while often retaining the responsibility of design, assembly, marketing and wholesale distribution.[16] One reason for such specialization is that the returns to scale and scope may vary substantially between the different stages. The (minimum) efficient scale in milling and dairy production, for example, is many times larger than the (minimum) efficient scale in farming. Similarly, a food retailer will want to exploit economies of scope by selling a wide variety of food items, such as reindeer meet and pineapples, while it makes little sense that the same farmer tends reindeers and grows pineapples. Such mis-matches between different stages of production suggest that a firm that

[15] In fact, different stages of the production chain are often seen as different *industries*.
[16] See Milgrom and Roberts, 1993. For an interesting account of recent developments in the mobile telephone industry, in particular concerning vertical specialisation, see The Economist, April 29, 2004.

tries to span the whole production chain will not be successful. Instead, one would expect to see firms specialising in one stage of the production chain only, or in a few adjacent stages where returns to scale and scope are comparable.

If returns to scale are substantial in one stage – perhaps big enough to make that stage a natural monopoly – while return to scale are much smaller in other stages, then a bottleneck problem is likely to exist. A lack of competition in one production stage will propagate into the other production stages.

There are different policy responses to this problem. One possibility is to choose a policy that suits one of the stages well. For example, ignoring the competitive problems in the bottleneck, one may opt for a laissez fairy policy or a policy of competition in infrastructure. Alternatively, the existence of a natural monopoly in the value chain may motivate that the whole of the industry becomes regulated or government owned. A final policy alternative is to adapt regulation to the particular characteristics of each successive stage, for example by introducing access regulation for the bottleneck stage, while relying on competition in other stages. The following two subsections will deal with the bottleneck problem as such, as well as with possible regulatory responses.

3.1 The bottleneck problem

The previous section discussed different sources of scale economies. The most important dividing line can be found between supply (or cost) side economies and demand side economies (or network effects). The former category is related to the production technology, while the latter is related to the intended use of the product and the benefit that comes from interaction between different consumers – for example when one person makes a phone call to another person.

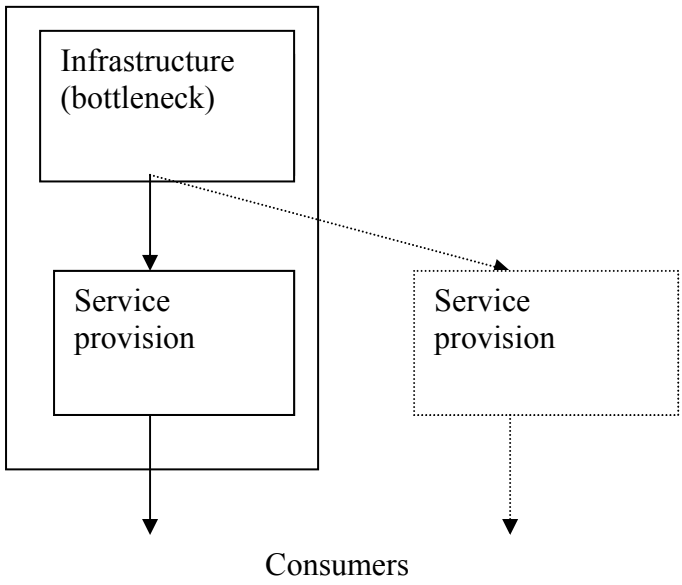
Related to this, there are two perspectives on network industries. One starts from the observation that there are economies of scale (or density) on the supply side in industries with a geographically dispersed infrastructure. It is more costly to duplicate, e.g., a railroad network, a telecom network or a network for transmission and distribution of electricity, than to use a single network more intensely. The other perspective on network industries starts from the observation that there are economies of scale (also) on the demand side, in industries that transport people or goods, or transmit information, between different geographical locations.

As argued above, telecom regulation (the requirement to interconnect) makes it likely that network effects between firms will be realised. Hence, network effects are not likely to be the major cause of bottleneck problems in the telecom industry. However, supply-side economies of scale may result in bottleneck problems, for the following reason. If there are substantial returns to scale in one production stage, then the efficient industry configuration will be one where a single firm controls this stage. A rival that tries to compete head-on is likely to incur losses and, predicting this outcome, the rival will be unwilling to engage in this kind of competition in the first place. Therefore, the firm that controls the critical production stage may be restrained neither by actual competition, nor by potential competition in that stage. Furthermore, the market power enjoyed in this stage may be leveraged into upstream or downstream production stages. In addition, if all other stages are competitive, the firm that controls one stage of the production will be able to earn approximately the same monopoly

profit as that attainable by a vertically integrated monopoly. In effect, the firm will be controlling a gateway position or, in other words, there will be a bottleneck problem in this industry.

Network industries, such as telecom, often have a production structure with several vertically related production stages. Because economies of scale and scope vary between the stages, competition is more viable in some stages than in others. This is often illustrated as in Figure 4, where only one firm can be active in the upstream infrastructure market, the bottleneck, while several firms can be active in the downstream market for service provision. For example, the upstream market can be establishing and maintaining a local telecom network. The downstream market can then be the telecom services market.

Figure 4. The bottleneck problem



By assumption, there will be market power in the bottleneck stage, which, in itself, gives rise to a number of negative consequences. (See section 2.5.) However, control of the bottleneck can give rise to market power also in the potentially competitive downstream market. In many instances, turnover in the bottleneck stage is relatively small, relative to total turnover. For example, a third of total revenues from fixed telephony in Sweden come from fixed fees (subscription fees).[17][18]

The idea that control over one monopoly stage in a succession of otherwise competitive stages gives control over all stages and hence yields the same profit as control over all stages would, has been recognised at least since the 1950s. This is sometimes known as the "law of one profit". Similarly, it has long been recognised that it may be more efficient to have one vertically integrated monopoly, than to have a succession of monopolies (Spengler, 1950). The reason is that an industry configuration with successive monopolists (or, more generally, a succession of firms with market power) will lead to "double marginalisation".[19] Hence, depending on one's point of reference, the law of one profit can be seen as bad or good. The bottleneck monopolist *will* be able to extract a monopoly profit and final prices will be as high as if there were a vertically integrated monopolist. On the other hand, prices would be even higher if there were a succession of firms with market power along the production chain.

More recent research has shown that a vertically integrated monopoly sometimes is preferable to a market structure where one stage is monopolised and the other stage is competitive, but also that the opposite may be true. Sometimes a firm that controls one stage may foreclose

[17] PTS, 2003. Arguably, at least some of the interconnection fees could also be seen as payments for the bottleneck controlled by the incumbent operator, TeliaSonera.
[18] For comparison, airport costs only constitute about a tenth of the total costs for air travel and the turnover of the central payment systems (e.g., the central giro and card transaction systems) is only a small fraction of total turnover in the retail banking market. Excluding taxes, the distribution and transmission costs account for more than half of the production costs for the electricity bought by a typical household, while this share is lower for large industrial consumers. See Bergman, 2002.
[19] That is, each firm in the value chain will add a margin on top of its marginal cost. This may result in a final price that is even higher than the monopoly price.

rival firms in the potentially competitive stages, to the detriment of efficiency (Tirole, 1988, ch. 4).[20]

3.2 Dealing with the bottleneck problem

The regulator – or government – can influence the functioning of markets with bottleneck problems in several ways. It can regulate, it can exercise control through ownership of (part of) the industry and it can influence the degree of horizontal and vertical integration within the industry. Seven main models have been used for bottleneck control, four of which can be seen as government control models and three of which can be seen as ownership structures that facilitate bottleneck control.[21]

The four control models are regulation of consumer prices, access regulation, monopoly franchising and government ownership (or hierarchy) control. The three ownership structures are vertical separation, horizontal separation and infrastructural clubs.[22] In addition, a laissez faire policy is of course possible. Under such a policy, the industry would be allowed to integrate vertically and horizontally and the government would exert no control on prices. If the bottleneck problem is sufficiently severe, this can of course result in an industry-wide monopoly and monopoly pricing. Alternatively, under more favourable circumstances, the outcome may be competition in infrastructure.

A fifth model of control could perhaps be added: government-induced competition in infrastructure. A direct way to achieve this would be to subsidise entrant owners of infrastructure, although this may be illegal under EU's rules for competition and state support. Indirectly, a similar result could perhaps be achieved through the design of access regulation. For example, initially favourable conditions for access can be used to stimulate entry into service provision. Subsequently, the access regime could become successively less favourable for the entrants, so as to induce investment in infrastructure. (See also section 5.3.)

It should be emphasised that the seven (or eight) models discussed below are broad categories. Within each of them, there are numerous variants and the choice between these variants can be immensely important. For example, a multitude of different methods can be used for regulation – including price caps, rate-of-return regulation, Ramsey pricing and Efficient Component pricing – with very different consequences for the functioning of the market. Furthermore, different models can be used in different segments or stages of the (telecom) market. In fact, many of the models can be used in combination also in the *same* market segment. For example, vertical separation is often used in combination with access regulation or government ownership. Finally, the distinction between the different models is not always apparent. A regulation of the monthly subscription fee for fixed telephone can be seen either as a (consumer) price regulation or as access regulation.

Regulation and control

Price regulation

Price regulation is perhaps the most typical model for bottleneck control. In Sweden, it has been used in the airline and the taxi industries, as well as in banking (rent regulation) and for

[20] See also Bergman, 2002, p. 123.
[21] Bergman (2002) identifies only six models. The seventh model, included in this text, is monopoly franchising.
[22] Government ownership could alternatively be seen as an ownership model. Here, the view is taken that government ownership is primarily a model of control.

electric utilities. In the U.S., it was the predominant model for controlling market power before deregulation; it was used in, e.g., the airline industry, for electric utilities, in the telecom and rail industries, and it is still being used in the taxi industry (Bergman, 2002).

There are many ways to regulate price. One important distinction is that between cost plus and price caps (Armstrong *et al*, 1994). Under cost plus, the firm's costs are compensated – which naturally gives it weak incentives to control costs. This is true for all types of costs, but a particular problem has been identified for investment expenses. Because of the long-lived nature of investments, the costs incurred by the investing firm must be spread over a number of years. In practice, this is often done by setting an upper limit for the allowed return on capital (or accounting cost of capital). This method is known as rate-of-return regulation, a special form of cost plus regulation. As long as the allowed return is higher than the true cost of capital, this gives rise to the Averch-Johnson effect: the firm will over-invest in order to increase its capital base (Averch and Johnson, 1962).

Under price caps, an upper price limit is set, either for each individual product or for a basket of products. Since the price ceiling is fixed, any cost savings that the firm can make will translate into higher profits. Hence, this type of regulation gives the firm strong incentives to control costs. However, any unexpected cost increases will also fall on the firm, which increases the firm's exposure to risk. In order to shoulder this additional risk, the firm must be compensated through a higher expected profit margin than it would need in a cost-plus contract (where much of the risk is born by the buyer). Given that firms are more risk averse than the buyer (perhaps the government or the collective of all consumers), this risk transfer is in itself inefficient. It follows that the optimal regulation can then be seen as a trade-off between an efficient risk allocation and good incentives for cost control (the "incentive – rent extraction trade-off"; see Laffont and Tirole, 1993). If too much risk is shifted to the firm, the expected price-cost margin will be too high; if the firm carries to little cost risks, it will have weak incentives to hold down costs.

A more complex scheme, in terms of the informational requirements, is Ramsey pricing. In order to set Ramsey prices, the regulator must have information both on demand. On the positive side, Ramsey prices achieve the optimal solution, given that government does not want to subsidise the industry. The main focus of Ramsey pricing is the price *structure*, while price caps focus on the price *level.* It is possible to combine these two, i.e., to use Ramsey pricing *and* price caps (possibly "global" price caps; see Section 5.3) so as to achieve both an efficient price structure and to set a ceiling for the price level.

Access regulation

Access regulation concentrates regulation to the bottleneck stage, under the assumption that if several firms are given access to the bottleneck, there will be effective competition in the other, potentially competitive, stages. This model has several advantages: it minimises the extent of the regulation and maximises the extent of competition. It also reduces the informational problem, both since the regulator only needs to estimate the costs in one production stage and since it will often be able to rely on the competitors' knowledge of the industry when dealing with the firm controlling the bottleneck.

There are some disadvantages with access regulation, however. The firm controlling the bottleneck will often have incentives to favour its own operations in the competitive stages. This can be achieved by inflating costs in the non-competitive stage, for example by

allocating costs from the competitive stages to the bottleneck stage (so-called cross subsidisation), or by reducing the quality of bottleneck services (infrastructural services) provided to the rivals. The latter possibility, in particular, necessitates a multi-dimensional regulation: a large number of quality aspects may need to be regulated, which makes the informational problem more severe (Laffont and Tirole, 1993, ch. 4).

Furthermore, price regulation of infrastructural services highlights the risk of regulation distorting the incentives for investments. If the regulation is too strict, the dominant network owner may choose not to invest, since much of the benefits will accrue to its rivals. At the same time, the rivals will have weak incentives to invest if the access regulation makes it favourable for them to rely on the dominant firm's infrastructure (Laffont and Tirole, 1993, ch. 1.9; Laffont and Tirole, 2000, pp. 137-139; see also section 4.3 below).

When it comes to the setting of prices, much the same methods can be used as when regulating consumer prices. However, Efficient Component Pricing (ECP) deserves a special mentioning. The main idea behind ECP is to stimulate efficient entry into competitive stages, while maintaining incentives for investment in the bottleneck. This is achieved by setting access prices equal to final prices (which may, potentially, be at the monopoly level) minus the dominant firm's avoidable costs in the competitive stage. With such (relatively high) access prices, it will be in the incumbent firm's interest to stimulate entry, while its investment incentives remain undistorted. Similarly, potential entrants will only have incentives to enter if they are more efficient than the incumbent. (ECP is discussed further in section 4.4.)

The public utility model

Public ownership of the whole or much of the network industry is the traditional European and Swedish model for controlling market power in industries with bottlenecks. In Sweden, it has been used in most network industries, including telecom, with banking being an exception. The public utility can be vertically integrated to include both the bottleneck stage and the potentially competitive stages, as was the case with Televerket (telecom), Postverket (postal services), SJ (rail) and Vattenfall (electricity). It can also be confined to the bottleneck stage, as has always been the case with the Civil Aviation Authority, and as is nowadays the case with Banverket (railway tracks) and Svenska Kraftnät (high-voltage electricity gridlines). Other examples are Terracom and Stokats

In public utilities, the central government can prevent market power from being exerted through its direct ownership control. Direct ownership also gives government flexibility to respond to changes within the industry, e.g., new technology or drastic changes in relative prices. The main disadvantage is that state ownership, in particular when combined with a monopoly position, does not give strong incentives for cost control.[23]

[23] For a theoretical analysis, see Laffont and Tirole, 1993, ch. 17.1; for references to the empirical literature, see Liu, 2001. Both the empirical and the theoretical literature suggests that private firms are, or can be expected to be, somewhat more efficient than state owned firms.

Monopoly franchises[24]

An alternative regulatory approach for a market that is not big enough to support regular competition is to auction a monopoly franchise. The firm that offers to provide services (of a specified minimum quality) for the lowest price will be awarded the monopoly franchise. In this way, a market mechanism can be used to drive down the price, even though the actual production will be undertaken by one firm only. At the same time, the winning firm's monopoly position may result in substantial scale economies. Ideally, competition *for* the market will deliver all the benefits associated with competition *in* the market, in addition to the benefits that can be derived from maximum scale economies. Relative to traditional regulation, an auction potentially gives the firms correct incentives for cost efficiency. In other words, the franchise auction *provides* information, while traditional regulation *requires* information. An additional benefit is that franchise contracts are mutually binding and, therefore, may give the franchisee a greater degree of certainty than a regulated firm would have. For the duration of the franchise contract, the terms are fixed, while regulation can be changed; this is the source of regulatory risk (see Section 4.2).

Unfortunately, there are some disadvantages with monopoly franchises. Short franchise tenures may distort the investment incentives. On the other hand, if the franchise tenure is too long, there are obvious risks that the franchisee tries to exploit its monopoly position, for example by degrading quality or manipulating the price structure so as to accomplish a de facto price increase. Hence, a regulation may still be necessary, with all the problems associated with regulation. In fact, a monopoly franchise is not likely to be a good solution for bottlenecks related to long-lived infrastructural assets. Finally, a monopoly franchise, at least in its simple form, means that there will be no competition in the potentially competitive stages.

Ownership structure

Vertical separation

Vertical separation has increasingly been advocated as a means of achieving efficiency in deregulation (OECD, 2001). As has already been mentioned, vertical separation has been used for SJ and Vattenfall, and has always been the norm for the Swedish Civil Aviation Authority. This method has also been used in other countries, e.g., for the former U.S. telecom monopoly, AT&T, and for British Rail.

Vertical separation does not in itself address the problem to which control over the bottleneck gives rise. However, it makes it easier to use other methods for controlling market power, notably access regulation and government ownership.

Access regulation is easier to implement over a vertically separated bottleneck owner for at least two reasons. The costs of the regulated firm accrue only in the bottleneck stage; hence there is no need to make assumptions on how to allocate common costs. In addition, a firm (or public utility) with activities only in the bottleneck stage should have no reason to discriminate between different firms in the potentially competitive stages. This makes it more likely that regulation can be one-dimensional, i.e., that access regulation will not have to specify the *quality* of the service provided, as would often be the case under vertical integration.

[24] See Viscusi *et al*. (2000) for a discussion of monopoly franchises.

Naturally, a disadvantage of vertical separation is that vertical synergies cannot be exploited.

Horizontal separation

Just as for vertical separation, the main advantage of horizontal separation is that it facilitates (price or access) regulation. Horizontal separation means that several firms will be active in the bottleneck stage, instead of just one, which increases the amount of information available for the regulator. In particular, benchmarking between the firms becomes possible.

Vertical separation has been used in electricity distribution in Sweden and elsewhere. It has also been used, e.g., for local and regional telecom services in the U.S. (the "Baby Bells") and for rail operation in the UK, Brazil and Mexico (Bergman, 2002).

If there are large returns to scale, vertical separation comes at a cost. Sometimes, however, the horizontal returns to scale may be smaller than the vertical synergies. At least, it appears likely that the cost of splitting up a network infrastructure into several geographically separated parts is often smaller than the cost of duplicating the infrastructure. (Cf. the discussion of returns to density in subsection 2.3.)

In particular circumstances, horizontal separation can be expected to give firms incentives to negotiate reciprocal access at relatively low rates. If these incentives are sufficiently strong, access regulation may not be necessary (OECD, 2001).

Note, however, that an important determinant of the outcome is whether there will be competition for customers or not (Cf. Table 1). If there is no competition for customers, the situation is like that for international interconnection. Then, if the firms (or countries) set their interconnection fees independently, there will be double marginalization, while if the fees are set cooperatively, they are likely to be set much closer to the efficient level. (Armstrong, 2002.)

If, on the other hand, there *is* competition for the customers, the situation resembles a competitive mobile telephony market.[25] However, in this setting it is likely that a horizontal separation of the bottleneck is not a feasible alternative, because of the scale economies. Conversely, if it is possible to have multiple owners of competing bottleneck infrastructure, then there could not have been a serious bottleneck problem in the first place.

Infrastructural clubs

The members of an infrastructural club are firms active in the competitive stage of a network industry, which jointly own the infrastructure. The model can be seen as an intermediate between vertical separation and vertical integration. Under favourable circumstances, infrastructural clubs are self-regulating. The firms have an incentive to keep costs low and, normally, they will have equal access to the infrastructure. In addition, vertical synergies can be exploited, at least to some extent.

However, although large firms will often be accepted into infrastructural clubs, small firms may face difficulties when seeking membership (see Katz and Shapiro, 1985, for an analysis).

[25] For an analysis of the incentives to set interconnection fees in such markets, and for the risk that multilateral interchange fees are used to coordinate retail prices, see Laffont and Tirole (2000) and Armstrong (2002).

In addition, if the infrastructural club has a monopoly, it may be used to coordinate pricing in the competitive stage.[26] For these reasons, infrastructural clubs are most likely to function (without regulation) when there is more than one competing infrastructure. This requires that most of the returns to scale are exhausted at volumes less than total industry output, although they may be large at the level of an individual firm.

Infrastructural clubs are common within the banking industry, in particular for payment systems, and are also used for the ticket reservation systems of the airlines (CRSs, computerized reservation system), taxi switches *et cetera*. Recently, infrastructural clubs have been set up for 3G mobile telephony infrastructure: one joint venture between TeliaSonera and Tele2 and one between Vodafone and Hi3G (or "3") (and, initially, Orange).

4. Short-run and long-run competition

The nature of the regulatory regime will determine the firms' incentives to compete, in the short run, as well as in the long run. Under consumer-price regulation, as well as under government ownership of the network industry, government retains much of the responsibility for the industry's development, both in the short run and in the long run. In the case of government ownership of the whole industry or, under vertical separation, government ownership of the bottleneck infrastructure, this obviously implies that (an arm of) government has to make investment decisions. However, consumer-price regulation can also require that government at least retains a veto right against new investments, because of the Averch-Johnson effect, as will be discussed below. This chapter will *not* discuss what investment criteria to use in such situations. I will, however, discuss how regulation affects the firms' incentives to invest and in what circumstances investment decisions can safely be left to the market – and in what circumstances they cannot.

If a regime with an infrastructural club is found to be the most efficient solution of the bottleneck problem, then the responsibility for the long-run evolution of the industry, including investment decisions, is delegated to the industry. The same is of course true for an unregulated (competitive or monopolised) industry, while a regime based on infrastructural-access can be seen as an intermediate solution. In the latter case, investment decisions are typically left to the firms, but the responsibility to create a good incentive structure will fall on the government (or the regulator).

The specifics of the access rules will matter greatly. For example, if the access price is set too low, neither the owner of the infrastructure, nor the downstream competitors will have incentives to make investments. On the other hand, given the existing infrastructure, short-run competition will be intense when the access price is low. Efficient-component pricing, briefly discussed in the previous section and further discussed below, is designed to provide strong investment incentives, but may fail to trigger sufficiently strong short-run competition. In fact, even in the long run, competition may be too weak under ECP: if there will be no investments by the potential entrants, despite correct investment incentives, then monopoly prices may prevail.

[26] E.g., by raising the price for infrastructural services to the monopoly level and then distributing the accruing profit between the club's members. Cf. the literature on patent pools, e.g., Lerner and Tirole, 2004.

4.1 Price regulation and long-run incentives

Above (section 3.2), two types of price regulations were briefly discussed: cost-based price regulation and price caps. If the regulator has perfect (or at least good) information on the industry's technology and the corresponding cost function, then cost-based regulation may be the best regulatory choice. The regulator can set the price equal to or slightly above costs, to the benefit of consumers. Another benefit of cost-based regulation is that it will reduce the regulated firms' risks. If the firms' risk aversion is high relative to the intrinsic risks of the industry, this may be a substantial benefit.

On the other hand, in many situations it is reasonable to assume that the regulator has less information on technology and costs than do the firms. In practice, the regulator will typically have to base the regulated price on accounting data provided by the regulated firms. This gives the firms incentives to inflate costs or, more or less equivalently, the firms will not have correct incentives to reduce costs. If there is a cost increase, it will just be passed on to consumers, since an increase in measured costs will raise the price ceiling. This, in turn, allows the regulated firms to be inefficient (so-called X-inefficiency) and to pay their employees well.

A possible way to achieve both low prices *and* correct incentives to reduce costs is to set a price cap at a reasonable level. This will ensure that consumers can benefit from a reasonably low price, while it will make the firms the "residual claimants" of any cost reductions. If the firms can increase their efficiency and reduce costs, this will not affect the price caps. Instead, cost reductions will increase the firms' profits. Just as in competitive markets, the firms' will have strong incentives to be efficient. Even if unexpected efficiency improvements will not benefit consumers, this is a benefit for society.

Of course, the above scenario is based on two assumptions: that the regulator can make a reasonable estimate of the firms' costs and that the cost risks are not too large. If the regulator cannot estimate costs, then the price cap may be set too high, resulting in too high profits and too little consumer surplus; or too low, resulting in negative profits and possibly exit from the industry or at least under-investment. If risks are very high, then the firms must be offered a high risk premium (i.e., prices must be high), in order to induce them to bear that risk.

However, there is an additional problem, which is related to the time dynamics. Due to, i.a., technological progress and trends in the costs of inputs, the cost level is not set once and for all. In the telecom industry, in particular, costs have been falling due to the technological developments. Hence, if a price cap is fixed once and for all, the profit of the producers will grow over time, or, if there is free entry, there will be excessive entry. The consumers, on the other hand, will have to pay a price that is unnecessarily high, with a price-cost margin that increases with time.

If this process is perfectly predictable, the problem can be overcome by incorporating a term that accounts for productivity gains, a so-called RPI-X scheme. That is, the price cap is indexed to a relevant price index, such as the retail-price index (RPI), but with less-than-total compensation for inflation. If the inflation is 5 % and X is 3 %, then the price cap only rises with 2 %.

However, the technological progress and other factors that influence the cost structure cannot be predicted perfectly. In consequence, the price cap will eventually become out of line with the evolution of best practice. Either the price cap will be too generous, resulting in excessively high profits and consumer prices, or it will be too strict, resulting in negative profits, lack of investments and, eventually, exit from the industry. Therefore, the price cap will eventually have to be re-aligned with the industry's cost structure.

The problem with re-aligning the price cap is that this re-introduces many of the problems with cost-based price regulation. If the re-alignment is made periodically, the firms will have an incentive to inflate costs towards the end of the period, so as to raise next period's price cap. If the re-alignment is made when costs have deviated sufficiently from the price cap, the firms will have an incentive *not* to reduce costs below the threshold that triggers re-alignment. Despite these effects, re-alignments will have to be made. Consequently, a price cap is not fundamentally different from a cost-based price regulation – it is just a cost-based price regulation with relatively long lags between price revisions. If a cost-based price is revised every year, a price-cap schedule is perhaps revised every five years.

4.2 Investments and competition, regulatory risk

The discussion in the previous section illustrated some of the additional complexities regulation has to accommodate due to the time dimension – even before investments are introduced in the picture. In practice, good incentives for investments are fundamental for an efficient long-run competition in telecom and in other network markets.

Firms need to invest, in physical capital as well as in R&D, marketing and other intangible assets, in order to produce efficiently. From a social point-of-view, investments are necessary, in order for efficient production. From the firms' point-of-view, however, investments will only be made if their expected returns are positive (or, more accurately, higher than the cost of capital).

A major complicating factor is the *non-reversible* nature of many investments. If an investment is made, part of the cost is often *sunk*. For example, a telecom operator's investment in a high-speed access network may be very costly but if, for some reason, the operator tries to sell that network, the second-hand market value can be much lower. In fact, unless another operator wants to use the network, it is probably completely worthless, since the main cost of building a network is the cost of digging. For other types of investments, a larger fraction of the investment cost can be recovered. For example, an airplane or a ship may be a significant investment, but since it can easily be moved, it will be valuable for other firms and its second-hand value may therefore be close to the initial price. However, as everyone who has tried to sell an almost new car knows, there may be an element of sunk costs even in the most liquid assets.

The existence of sunk costs makes regulation more complex. Once a sunk cost has been incurred, the firm's incentives change. Before the investment is made, the relevant cost for the firm's decisions is the whole investment cost, but *after* the investment has been made, only the opportunity cost is relevant – and sunk costs are not part of the opportunity cost. Let's return to the example of a newly built high-speed access network!

Once the network has been built, the firm will remain active in the market as long as it is able to cover its variable costs, which do *not* include sunk investment costs.[27] Hence, a regulator can impose a very low access price, without inducing the firm to exit the market. Seen in isolation, such a policy will actually be welfare improving – since welfare is maximized if the access price is set equal to the marginal cost of providing access. An ambitious regulator may therefore be tempted to set very low prices. On the other hand, if the regulated firm is able to predict this outcome – or if it thinks that there is a risk for such an outcome – then the firm will be much less inclined to make investments in the first place.

In the context of regulated markets, the above scenario is known as *regulatory risk*. A similar phenomenon may arise also in the relation *between* firms that make so-called relation-specific investments. For example, a coal-mining company and a power producer with a power plant located at the mouth of the mine are mutually dependent – and are each in a position to take advantage of the other. If the mining company invests in the mine, the power producer may try to exploit the sunk-cost nature of the investment and reduce the price paid for the coal. Conversely, if the power producer invests in a new boiler, the mining company may try to *raise* the price of coal. [28]

There are different ways to resolve these types of ex-post opportunism. One possibility is to avoid making non-reversible investments, for example by hiring physical capital, rather than by buying it. For example, an airline company may lease airplanes, rather than buy them, and a shipping company may lease ships. Such a solution, however, would not be very helpful in the context of investments in telephone access networks, since it would only shift the regulatory risk from the operator to the leasing company. The mining company and the power producer would not benefit much from transferring the investment risk to a third party either, but there are two other methods that can be used: vertical integration and long-term contracting. In a regulatory setting, the closest correspondences would be government ownership (integration of the operator and the regulator) or long-term *commitment* by the regulator not to exploit the sunk-cost nature of the investments.

Given that the government does not want to become the owner of the operation (i.e., the public-utility model discussed in section 3.2) and given that leasing the capital will not resolve the problem, the regulator will need to ensure sound incentives for investments by establishing a predictable regulatory regime. It is not enough to declare that the regulator will not engage in ex-post opportunism, since such a statement would be much too vague. Since measuring costs is not an exact science, there will be many opportunities for the regulator and the regulated firms to make different interpretations on how large cots actually are. The root of the problem is of course that there is no single correct way to allocate common costs between different activities and no single way to allocate investment costs over time.

Hence, establishing a predictable regime includes establishing principles for how investment costs should be calculated (backward-looking or forward-looking, the allocation of common costs et cetera), what return on investments is acceptable, over how long period different investments should be amortized and *which* investments to include in the cost base.

The last point alludes to the Averch-Johnson effect mentioned earlier. If the regulation is designed in such a way that investments are profitable for the firm, even if they are not

[27] Even if the firm itself goes bankrupt, its assets are likely to remain in the market, although with a new owner.
[28] Paul Joskow has published a number of studies on contractual relations between coal mines and mine-mouth power plants. See, e.g., Joskow (1987).

profitable from a welfare-point of view, then the regulator may need to have some *de facto* veto power concerning investments (even if there is no such formal authority). If this were not the case, the regulated firm may be tempted to inflate its capital base, in order to increase costs and thereby prices. Assume, for example, that a regulated firm's cost of capital is 8 %, while the allowed rate of return on capital is 10 %. Then an excessive investment of 1 billion would increase the firm's accounting costs with 100 million, while its true cost increase would be no more than 80 million. Assuming that the investment is completely useless, the resulting annual loss for the consumers would be 100 million, firm profit would increase with 20 million, and the net annual welfare loss would be 80 million.[29]

In order to reduce the risks of excessive investments, under US regulatory regimes, the regulated firm was typically allowed to include in the cost base only such investments as were "used and useful".[30] Even if the regulator did not have the formal authority to block investments, that was the *de facto* result of the principle. (See further Laffont and Tirole, 2000, sect. 4.4.1.4 and 4.4.2.)

4.3 The "stranded-assets argument"

The above section has illustrated the importance of establishing a regulatory regime that avoids ex-post opportunism. The regulated firm that makes an investment must be allowed to earn a fair return on its investments, which requires that sound principles are established for calculating investment costs, and the firm must be made *confident* that this will happen, which requires a certain degree of regulatory commitment. However, this is not the end of the story. The firm may also have to be compensated for the possibility that its investments *could* have failed, even if they did not.

The importance of the so-called *stranded-assets argument* is apparent in the context of pharmaceuticals. In most countries, the prices of pharmaceuticals are regulated, with prices set at a level that not only allows production costs to be recovered, but also the R&D costs needed to develop new medical substances. However, in order to develop one commercially successful medical product, a large number of attempts has to be made. Assume that a pharmaceutical firm makes some preliminary investigations into 1000 substances, which may be intended for, say, 10 different medical uses. Assume also that 100 of these appear to have some positive effect in the initial experiments, perhaps on animals. These will then be tried for positive effects and for negative side effects on humans, in so-called phase I through III studies. Assume, finally, that only a handful of the substances eventually are approved by the medical product agencies and marketed as drugs – and that only one of them is commercially successful.

Now, the regulated price of that one successful product must be high enough to allow the firm to recover its R&D investments on *all* 1000 substances. Naturally, the apparent price-cost margin on that substance will often have to be high. A possible solution would be to base price regulation on the *firm's* rate of return, rather than on the costs of a single medical product. However, there are important problems associated with this approach as well. First, internal-transfer prices within multinational pharmaceutical firms may make such regulation

[29] This is a simplified version of the Averch-Johnson effect, assuming, for example, that there are no substitution possibilities between capital and other factors of production.
[30] See Armstrong *et al*., 1994, p. 87

ineffective. Second, and more fundamentally, the original research into the 1000 substances may not have been undertaken within one firm. Possibly, 1000 small independent firms could be financed through the venture capital market, even though the investors accurately predict that only one out of a thousand products will eventually be successful. Then that single successful firm would have to be compensated for the losses incurred by the other 999 firms – resulting in extremely high returns for the lucky investors who happened to pick the right investment alternative.

A similar line of reasoning may be applied to telecom and other regulated network industries, although the risks here are perhaps not as extreme as in the above example. If a firm undertakes risky investments in infrastructure, a regulated access price should not only reflect a fair rate of return on the actual investments, it should also compensate for the risk that the investment turned out to be unsuccessful. In other words, a cost-based access-price regulation should not only be based on an ex post cost measurement, but also on an estimate on the ex ante investment risks.[31]

Naturally, establishing the ex ante risks may be extremely difficult. However, this line of reasoning suggests that a smaller profit margin is justified for networks that were established during a period of monopoly protection, relative to networks that were built in competitive markets. In markets where licenses are auctioned, one possible method for estimating ex ante uncertainty is to study the bid spreads.

Another aspect of the stranded-assets argument is that it may be reasonable to compensate firms for assets that are not used, or unprofitable. Firm sometimes make investments that turn out to be mistakes. Two possible examples are high-speed access networks that quickly become obsolete, because of new wireless technology, or nuclear power plants that are so costly to build that the associated investment costs can never be recovered. To some extent, it may be reasonable to allow the firm to recover these losses by raising the regulated prices of other services, such as telephone access fees in general or the electricity price in general. (The electricity example is more relevant for the US market, where electricity prices are still regulated in many states.) On the other hand, regulation cannot be designed so that firms are given a *carte blanche* for recovery of bad investments, at the expense of consumers. (See Laffont and Tirole, 2000, section 4.1.1, and the above discussion of the Averch-Johnson effect.)

4.4 Essential facilities

Sometimes, general competition rules will oblige dominant firms to provide access to vital infrastructure. Such obligations will of course be of relevance when evaluating the degree of realism in the "sunset vision" – i.e., in repealing sector-specific (telecom) regulation and just relying on general competition rules. Also, to the extent that access will be required under the competition rules, sector-specific rules for access will of course be superfluous.

[31] Bergman (2004). This distinction is not to be confused with that between ex post and ex ante regulation, i.e., whether the regulator should set the access price in advance or address excessive access prices in retrospect, possibly by requiring partial repayment by the owner of the infrastructure to the service providers. It is more closely related with the distinction between forward-looking and backward-looking cost calculations. In fact, the ex ante perspective advocated in the text above is most closely related to backward-looking cost calculation.

Only dominant firms, in the specific sense of competition law, will ever be required to provide access under the general competition rules – and even dominant firms will often *not* be required to provide access to competitors. The set of circumstances when such an obligation exists can be found in the so-called essential-facilities doctrine. According to this doctrine, a dominant firm will have an obligation to provide infrastructural access when a) access to the infrastructure is necessary in order to compete in a related market, b) the competitor is unable to build, acquire or maintain its own, alternative, infrastructure and c) when the owner of the infrastructure competes in the related market.[32] If there is an access requirement, access has to be provided at non-discriminatory terms. In particular, the terms offered must be non-discriminatory vis-à-vis those offered *within* the dominant firm. An interpretation is that access must be provided at cost. In practice, the permitted mark-up above, e.g., the LRIC price will typically be quite large.

Criterion b above – the competitors' inability to duplicate the facility – has been the subject of much debate. In the so-called *Bronner* case, the EC Court stated that in order to establish that duplication is not possible, it must be shown that even if a competitor held half of the market, it still would not be economically possible for that firm to build and maintain a second facility. Bergman, 2004, discusses this *"Bronner"* criterion" further. Although it provides a substantial degree of predictability, it contains a double risk of induced inefficiency. First, the criterion may impose an access obligation even in situations where this risks chilling investment incentives. Second, it may *not* impose access obligations in situations where there are no adverse effects on investments, but where duplication of infrastructural assets would simply not be cost effective.[33]

To conclude, general competition rules and the essential-facilities doctrine can be of some help for an entrant firm that seeks access to a dominant firm's infrastructure. However, the obligations imposed on the incumbent firm will be less strict, and will apply in fewer circumstances, than what we have become used to as the norm in telecom.

5. Competition in services or competition in infrastructure?

Access regimes are often designed with an ambition to achieve several partially conflicting goals simultaneously: to encourage efficient use of the infrastructure, to support policies for universal service, to allow the owner of the infrastructure to recover fixed network costs, to give incentives for cost reduction in the provision of infrastructure and to give incumbent and entrant network operators correct incentives to invest in new infrastructure.

If the only instrument available to the regulator is the access price, the regulator's task may in fact be impossible. Efficient use of the existing infrastructure suggests that access should be provided at marginal costs. However, this means that fixed network costs will not be covered and "cherry-picking" by the entrants (i.e., entry only in profitable markets) will perhaps drive the incumbent out of profitable markets, with the effect that universal service obligations cannot be sustained.

[32] This is a simplification; see Bergman, 2001 for details.
[33] To some degree, the first risk has been addressed in the recent *IMS* case.

There is, however, one instrument that (under ideal conditions) is potent enough to deliver efficiency in all dimensions simultaneously. That instrument is called competition. Although conditions are never ideal – certainly not in the telecom industry – a promising idea appears to be to use competition as much as possible, and regulation no more than necessary. In fact, this is the main rationale for the "deregulation" of the telecom industry and the increased reliance on access regulation, rather than (consumer) price regulation. The bottleneck is regulated, while competition in services is made possible by access regulation.

The scope of the bottleneck, however, is not given once and for all. Technological progress may transform a natural monopoly into a potentially quite competitive market. In addition, as argued above, it may sometimes be worthwhile to duplicate facilities that are natural monopolies. Hence, a natural extension of the policy of "deregulation" is to encourage facilities-based competition. According to Laffont and Tirole (2000, p.8 and 22), dismantling sector-specific regulation has often been seen as desirable, once competition has developed to the point where it can be maintained with the general competition rules.[34] In the process of launching the E-com directive, similar opinions have been expressed by the EU Commission and others.[35] A commonly held view is that facilities-based competition is a pre-condition for competition to evolve and, hence, a pre-condition for dismantling regulation. Extending the above line of argument, facilities-based competition is, in itself, seen as desirable. A key argument of the present paper, however, is that the benefits of facilities-based competition must be weighed against the costs of duplication. If the latter are too large, then access-based competition will be the better alternative.

There is a certain degree of inconsistency between the endorsement of the "Sunset vision" and the de facto emphasis on access regulation in the E-com directive. One way to reconcile these seemingly incompatible approaches to regulation is to use access regulation to create a "beachhead" (in terms of entrants' market shares, say), from which competition can evolve and *then* to lift regulation. More specifically, according to the "investment-ladder hypothesis", the regulator can foster facilities-based competition through the implementation of a carefully designed dynamic access regime. In the initial stages of competition, access is mandated at terms that are favourable for the entrant, in order for the entrant to build market shares and to counter the effect of the incumbent's huge first-mover advantage. However, as competition evolves and as the entrants' market shares increases, access regulation becomes less favourable for the entrants which. This gives the new firms progressively stronger incentives to invest in their own infrastructure.[36]

From a practical point of view, the regulator (or the policy maker) can choose between policies that promote competition at various levels in the value chain:[37]

- Pure service provision, for example long-distance telephony services, based on origination and termination access; dial-up internet-service provision

[34] See also Bergman *et al*, 1998.
[35] See Oldale and Padilla, forthcoming, for references and quotations.
[36] For a more elaborate treatment of this view, see Cave and Vogelsang 2003 for a critical comment, see Oldale and Padilla, forthcoming.
[37] Valletti (2003) outlines the three basic modes of entry: 1) *Facility-based competition*: Both incumbent and entrant build their own network and compete directly for customers. If there is access regulation, it will concern two-way access (for call termination); 2) *Local loop unbundling*: The entrant is able to lease the incumbent's access facilities. Regulation may concern line rental (one-way access) and call termination (two-way access); and 3) *Carrier selection*: Every call originates on the incumbent's network; regulation now concerns one-way access for both origination and termination.

- Resale entry, for example resale of the incumbent's fixed telephony subscription services and broadband services
- Mixed (or unbundled) entry; a combination of resale entry and facilities-based entry, whereby the entrant buys some "unbundled elements", i.e., specific infrastructural services, from the incumbent, while producing other infrastructural services in-house
- Facilities-based entry, for example competitive provision of fixed-line services or broadband services

One concern of the policy maker should be to encourage only efficient entry. There is one well-known policy rule that has been designed specifically to achieve that: the Efficient Component Pricing rule, or the ECP rule.

5.1 Efficient-component pricing

The ECP rule, or the Baumol-Willig rule as it sometimes is called, was proposed as a method for creating incentives for efficient entrants (only) to enter the market. According to the ECP, the access price should be set at a level that compensates the network owner both for the cost of providing access and for the loss of profit due to lost sales and the ensuing loss of mark-up on these units.

According to the rule, the access price should equal the incumbent's opportunity cost of providing access. The opportunity cost has two components: first, the marginal cost of providing access and, second, the incumbent's loss of the mark-up on the retail sales. Assume, for simplicity, that the total quantity demanded is unaffected by the activities of the entrant: if one more unit is provided by the entrant in the retail market, then the incumbent will sell exactly one unit less. Assume also that the marginal cost of providing access is c^1 and that the (incumbent's) marginal costs in the (competitive) retail stage is c^2, while the retail price is p. Then the first component of the incumbent's opportunity cost is c^1, the marginal cost of providing access, and the second component is $p- c^1- c^2$, the incumbent's loss of profit due to the loss of one unit of sales. Hence, the incumbent's total opportunity cost is $p- c^1- c^2+ c^1=p- c^2$.

In other words, under the simplifying assumption of an inelastic total demand, the ECP rule prescribes that the access price should equal the retail price minus the incumbent's marginal costs in the retail stage. The principle can be illustrated with a numerical example.

Assume that the incumbent has marginal costs 2 and 3, respectively, for providing access and in the subsequent retail stage. Assume also that the retail price is 6. According to the ECP rule, the access price should be 2 + (6-2-3) = 3, i.e., the per unit access cost plus the per unit profit.

If the entrant's and the incumbent's products are not perfect substitutes, the incumbent will not loose one unit of sales for each unit of sale achieved by the entrant. If the goods are not substitutes at all, there will in fact be no loss of sales at all. The ECP formula can be adjusted to account for this, but it then becomes slightly more complicated (Armstrong, 2002). The access price should now be equal to the sum of the marginal cost of access, which is still c^1, and the opportunity cost, which will now be $s(p-c^2)$, where s is the number of units of sale lost

for the incumbent for each unit of sale made by the entrant, and $p-c^2$ is the gross retail profit. Hence, the modified ECP says that the access price should equal $c^1+s(p-c^2)$.

The variable s can be assumed to lie in the [0,1] range. If the entrant's product is a perfect substitute for the incumbent's produce, then s=1 and we are back at the simple formula above. This may be a good approximation for the market for long-distance telephony: if one more phone call is made through the entrant's long-distance network, it is not unreasonable to think that one phone call less will be made through the incumbent's network.
If the products are not substitutes, then s=0, since increased sales by the entrant will not reduce the incumbent's sales at all. Then the ECP prescribes that access should be provided at marginal cost, i.e., c^1. This may be relevant for new services, which are not provided by the incumbent. In the general case, s will be between 0 and 1. Note that now the access price will differ between different types of entrants, depending on the substitutability between their products and those of the incumbent.

At the price prescribed by the ECP rule, the incumbent will earn the same profit from selling retail services itself or from providing access to the incumbent. (In the general case, with $s<1$, it is more correct to say that the incumbent's profit will not fall because of the entrant's activities.) Hence, the incumbent has no incentives to refuse to sell to the entrant and, in particular, it will not have incentives to degrade access quality. The entrant, on the other hand, will only have incentives to enter the market if it is more efficient than the incumbent. This avoids the possibility of inefficient entry, or "cherry picking". With a uniform cost-based access price, the entrants could otherwise be expected to enter those market segments where the retail price is high, relative to retail costs. This, in turn, could potentially be the result of a pricing scheme that tries to recover common costs with mark-ups that are high on those market segments where the price elasticity is low (i.e., Ramsey pricing) or of a universal service obligation imposed on the incumbent (see Hultkrantz' chapter on USO in this volume).

However, the ECP rule cannot resolve all issues. It will not provide the owner of the infrastructure with incentives for cost efficiency in the infrastructural stage. Also, the rule has only limited effects on the retail price. Given that the retail price was initially set at the monopoly level, it will remain close to that level. If the policy maker wishes to bring prices down to a competitive level, an ECP rule must be combined with policies for curbing the retail price, such as direct regulation at the retail level (Armstrong, 2002; Valletti, 2003). Furthermore, ECP may give the incumbent incentives to choose a technology that gives it low marginal costs in the retail stage, even if fixed costs increase so much that total costs increases. Finally, ECP may result in inefficient entry in the market for infrastructure. This may happen if the entrant's marginal cost in the infrastructural stage is higher than the incumbent's (higher than c^1), but lower than the incumbent's opportunity cost (lower than $p-c^2$).

The ECP has received a lot of criticism from regulators (see Laffont and Tirole, 2000, and Valletti, 2003). Interestingly, a version of the simple ECP has re-surfaced in recent applications of competition law, under the label "margin squeeze". (Armstrong, 2002, calls the simpler version of ECP, i.e., access price = $p-c^2$, the *margin rule*.) According to legal practice, a dominant telecom operator may be abusing its dominant position, i.e., violating Article 82 of the EU Treaty, if it only provides access at a cost that is so high that an entrant cannot reasonably compete. This has been interpreted as an access cost that is so high that the margin between it and the retail price is not big enough to cover the incumbent's marginal (or

incremental) costs in the retail stage, i.e., if $a>p-c^2$. Rewriting this as an equality, we see that the expression is identical to the (simple version of) the ECP rule.[38]

5.2 Efficient and inefficient bypass

The ECP rule focuses on efficient entry in service provision. As noted in the previous subsection, it may result in inefficient entry into the infrastructural market. However, it is conceivable that an alternative policy can be designed, such that what the ECP rule achieves in the retail market can be achieved also in the market for infrastructure.

Inefficient entry into the market for infrastructure is sometimes called *inefficient bypass*, the term I will use below, or *inefficient network duplication*. (Laffont and Tirole, 2000, section 3.3, reserve the first term for inefficient investments made by large customers and the second term for inefficient investments made by the incumbent's competitors. See also Armstrong, 2002.)

If the incumbent's access prices are high only because it exerts market power, then inefficient bypass would not be a problem. If the entrants are in fact more efficient, the threat of bypass will force the incumbent to lower prices, which will be good for efficiency.[39] (A lower access price, in turn, will make bypass investments less likely, but this will only be good for efficiency, if the provision of infrastructure is a natural monopoly.) If, on the other hand, the entrants are *not* efficient, then bypass is not efficient.

However, there are sometimes good reasons for the incumbent to keep access prices above costs in some segments. There may, for example, be fixed network costs that must be recovered by excess profits in some segment of the market. In economics jargon, there may be an *access deficit* that has to be covered by markups. If access is priced at costs and if the incumbent tries to cover the access deficit by markups in the services markets, then there will be inefficient entry in the services markets and the ensuing competition will drive the incumbent from the services market. (Cf. the section on the ECP rule.) If, instead, the incumbent marks up its access services, then there may be inefficient bypass. If a universal service obligation is imposed on the incumbent provider of infrastructure, the access-deficit problem will be aggravated. Now competitive access providers will have even stronger incentives to enter low-cost access markets, such as metropolitan areas. (Cf. Hultkrantz, this volume.)

Inefficient bypass has at least two negative consequences. First, the incumbent will not be able to recover the access deficit or fulfill its universal service obligations. Second, inefficient entry represents a socially wasteful duplication in a market characterized by natural monopoly.

5.3 Policies for efficient access competition

[38] For a discussion of recent cases under EC competition law, see Garzaniti and Liberatore (2004).

[39] On the contrary, there is a risk that bypass would not be a sufficient threat. If an entrant has to sink costs in order to enter the market for infrastructure, then the incumbent will likely be able to maintain prices in excess of the competitive level. See the literature on contestable markets.

At least three methods have been proposed to provide correct incentives for competitive investments in infrastructure: output taxes, global price caps and escalating access prices. These proposals will be discussed in turns.

Output taxes

Armstrong (2002) advocates the use of an output tax imposed on the entrants, in combination with cost-based access fees. Then the entrant will have correct incentives in the make-or-buy decision concerning infrastructure, while the output tax can be used to cover the incumbent's access deficit. As noted by Armstrong, this sounds like a radical and perhaps discriminatory idea: entrant telecom operators will have to pay a "tax" on their output and the tax revenues will go to the incumbent.

However, the output tax can equally well be seen as (and called) a fee that is used to meet universal service obligations. In fact, the output tax can be imposed on incumbents and entrants alike, with the proceeds going to the firm that meets the USO, possibly after a procurement process. Consequently, the output tax sometimes goes under the name USO fee. Laffont and Tirole (2000) call it excise tax or retail tax.

Global price caps

Laffont and Tirole (sect. 4.7) argue in favour of another scheme: a so-called global price cap. The idea is to subject the incumbent to an average price cap that applies to wholesale (infrastructural) services as well as to retail services. The task of the regulator will be to set the price cap, but subject to the price cap, all prices will be set by the incumbent. If the price cap as set at the appropriate level, the incumbent will have the power to extract enough profits to cover the access deficit and meet possible USOs, while at the same time the price cap will prevent it from setting prices higher than necessary. It will be in the incumbent's interest to set high markups in market segments with inelastic demand, whether retail or wholesale, and low markups in market segments with elastic demand. In fact, the incumbent would be using a variant of Ramsey pricing or, in other words, a socially optimal price structure, given that it has to cover its costs.

In principle, the same thing could be achieved by a regulator, if the regulator had the power to set retail prices as well as access prices. However, the regulator would face a monumental informational problem: just as with traditional Ramsey pricing, the regulator would need to know not only the cost structure, but also demand elasticities for all end products and intermediate services.

An additional benefit of the method is that it will not be in the incumbent's interest to discriminate against its competitors with non-price methods, since it will be just as legitimate for the incumbent to make a profit on access provision as on service provision. In fact, in this setting, "excluding buyers of interconnection services amounts to mutilating a potentially quite profitable activity".[40]

A potential problem with the global price cap, however, is that it would give the incumbent the possibility (if not the incentive) to instigating a price squeeze. Under the global price cap, the incumbent could simply reduce retail prices and increase access prices. Although this

[40] Laffont and Tirole (2000), p. 174.

would not be profitable in the short run, it may be profitable in a long-run perspective, when taking into account future revisions of the price cap, and the threat of facilities-based competition being established. Laffont and Tirole's solution to this problem is an active use of competition policy, which can be quite effective against price squeezes.

Escalating access prices

Cave and Vogelsang (2003) have suggested that access prices should rise over time, in order to induce competition in infrastructure and so as to reward investments. This may seem like a theoretical notion, without much relation to regulatory practice which, in effect, has historically led to *declining* access fees. However, the new E-com directive clearly suggests that the scope of the regulatory intervention should be adapted to the competitive pressure. If competition is absent, relatively stringent conditions (such as cost-based access provision) are imposed on the owner of the infrastructure. As competition becomes more intense, these access requirements become less stringent. Then, such measures as non-discrimination and transparency will be applied. Eventually, when the incumbent is no longer dominant, it will no longer have any particular obligations.

5.4 Competing infrastructural clubs and specialized access providers

Above, it has implicitly been assumed that the potential entrants into the infrastructural market are firms that consider bypass investments that would provide access for their own downstream operations. There are, however, other possibilities. One is that specialised access providers try to enter the market; another is that several downstream service providers jointly invest in infrastructure. In the latter case, a possible industry configuration would be two or three competing providers of infrastructure serving several firms active in the downstream market for services. In the Swedish mobile telephony market, for example, there will be two competing infrastructural clubs that provide services to four competing downstream service providers.

Competing infrastructural clubs can be seen as a compromise between facilities-based competition and access-based competition in service provision. There will be some duplication of assets, but less so than if each firm has to build its own infrastructure. At the same time, as argued in section 3.2, infrastructural clubs can potentially be self regulatory – i.e., the incentives can potentially be such that the firms compete *and* minimize costs in the infrastructural stage. Possibly, the incentives will be even better when there are competing clubs.

A disadvantage with infrastructural clubs is that they can provide an instrument for coordination between the downstream competitors. When the downstream competitors decide on the access price, they can potentially set the price in such a way that each of them individually has incentives to set high prices in the downstream market. Infrastructural clubs controlled by a few large firms may also be reluctant to accept small firms, which contributes little to the network effects, but which increases competition for customers. However, these concerns are likely to be less pronounced in a setting where two or three infrastructural clubs co-exist on the market. If one club rejects an applicant firm, the applicant may be accepted by the other club, reinforcing network effects within that club. This scenario is likely to make the clubs more willing to accept newcomers in the first place.

In a setting with two infrastructural clubs owned by four downstream competitors, Nordberg (2004) has recently shown that the coordination effect is also likely to be less problematic when the club is not in a monopoly position. In fact, the downstream owners may sometimes have incentives to set access prices *below* marginal costs. In particular, this will be the case when the downstream services provided by all four firms are close substitutes. If, instead, the services provided by firms associated with one of the infrastructural club are not good substitutes for the services provided by the two other firms, then the infrastructural clubs may indeed have incentives to set high prices. An example of the latter is perhaps competition between different generations of mobile telephony, or between two networks with large differences in coverage. An example of the former may be competition between two same-generation networks with much the same coverage.[41]

Specialised access providers present the regulator with a particular problem. In general, they will want the regulator to set *high* access prices. If the regulator sets low access prices, competitive access provision will not be profitable. On the other hand, if the regulator sets high access prices – or if there is no access regulation at all – then access provision will be more profitable. The presence of a specialised access provider will limit the incumbent's ability to foreclose entry. However, just as with entry by integrated firms, there is a risk of inefficient bypass.

6. A comparison of telecom with electricity and postal markets

Although the telecom industry, postal services and the electricity market were all "deregulated" at around the same time[42], quite different regulatory policies have been used for the three industries. To some extent, differentiated regulatory strategies will reflect differences in underlying technologies and market characteristics: the regulatory framework must be adapted to the particularities of the regulated industry. On the other hand, to some extent the differences will be the result of circumstances that are irrelevant from an efficiency point-of-view. It is likely that important lessons can be learnt from a comparative analysis of the regulatory histories of different industries.

In the telecom industry, relatively stringent access obligations have been laid on the incumbent (the owner of the network), but vertical integration has remained the norm. In electricity, transmission (the high-voltage "core" network) has been vertically separated in Sweden and some other countries, while distribution (the low-voltage "peripheral" network) has remained vertically integrated with generation and sales. In Sweden, electricity distribution has, in principle, been subject to the same kind of cost-based access regulation as the telecom networks, although the regulation has been ex post rather than ex ante and, in practice, the regulation has been less stringent.

In postal services, there have been virtually no access requirements at all (except access to boxes and postal codes). In the postal market, the key bottleneck facility appears to be the

[41] Nordberg's analysis is static. If multi-period competition is introduced in the analysis, the firms may be able to sustain collusion through other mechanisms, both within an infrastructural club and between them.
[42] In Sweden, telecom and postal services were deregulated in 1993, while electricity was deregulated in 1996. Within EU, telecom deregulation began in 1990, with an important reform in 1997, electricity deregulation began in 1999 and postal services have gradually been deregulated since 1997. See Bergman, 2002.

mailbox delivery system, i.e., the postmen and their routes. To some extent, sorting terminals and collection can also be seen as bottlenecks.

Focusing on Sweden, this means that the regulatory policy for the electricity market has been based on competition in services, just as in the telecom industry. The regulatory policy for postal services, on the other hand, has been one of facilities-based competition. As a consequence, competition in postal services has developed only in those market segments where demand is high enough to support two bottlenecks: large batches of mails for delivery in the major metropolitan areas. In these segments, the main new entrant's market share is approximately 30 %, but the incumbent operator, Posten, maintains a market share of approximately 95 % of all mails. Furthermore, because of the limited overall competitive pressure, consumer-price regulation has remained in place. Bergman (2002) and Andersson (2004) have proposed the introduction of some type of access regulation and Andersson also proposed the repeal of the price-cap for standard mails. In line with the theoretical arguments of the previous sections, the choice of a policy of facilities-based competition has made it possible to use "light-handed" regulation.

The electricity market offers some interesting comparisons with the telecom market. In telecom, the long-distance "core" network appears to be a well-functioning competitive market, due to relatively modest costs of duplication. In other words, facilities-based competition is already at hand. The peripheral parts of the telecom network, in contrast, are more costly to duplicate. Consequently, access regulation has focused on access to these. Although facilities-based competition in this segment would lead to costly duplication, it would also stimulate investments in new technology and it would perhaps make a more light-handed regulation possible.

In electricity, both core and peripheral parts of the network are natural monopolies, in the sense that duplication is not a realistic alternative. Consequently, facilities-based competition has not been much of an issue.[43] In electricity distribution, the two main concerns have been that relatively lax regulation has allowed too high prices and that there is too little maintenance investments, with the effect of interruptions, for example after heavy snowfall. The regulatory response has been to make regulation more stringent, i.e., to follow the regulatory model used for telecom. Since competitive investments are not envisaged and since there is no need for an extensive up-grading of the distribution network, incentives for investments has not been an important issue. This is in contrast to the situation in electricity transmission.

Different models have been used for regulating transmission within the European Union. In the Nordic countries, vertical separation has been combined with government ownership of the transmission network. In the UK, the vertically separated network is privately owned, while in continental Europe, vertical integration with power generation seems to be the norm. Under the "Nordic" model, high transmission prices is not a key issue, while in the UK and on the Continent, access prices have to be regulated in much the same way as (local) telecom access is regulated.

Helm (2001) argues that the British model of vertical separation and access regulation has been good at "sweating the existing assets" in the electricity sector, i.e., to achieve efficient use of the existing infrastructure, but that it has been much less successful in achieving

[43] Of course, electricity generation requires facilities and large investments. From the perspective of bottlenecks and competitive market segments, however, competition in generation can be seen as competition in services.

dynamic efficiency, i.e., the appropriate level of investment. He argues that neither access-based competition, nor facilities-based competition in combination with general competition law is likely to achieve dynamic efficiency.[44] Instead, a more coherent (and interventionistic) policy is required. One reason is that private investors will fail to recognize the asymmetric costs associated with non-optimal investments in infrastructure: excessive investments will be costly, but insufficient investments will be much more so, from a social point of view. In Helm's view this suggests a national policy that promotes reserve capacity in the networks.

The problem of insufficient investments in electricity transmission is not confined to the UK. The consensus view appears to be that relatively large investments in transmission capacity, in particular *between* countries, are necessary to reap the full benefits of market liberalisation. However, more cross-border transmission capacity would lead to more intense competition for customers in previously insulated national markets – and this is not in the interest of the incumbents which, in large parts of Europe, own the transmission network. Even in the Nordic countries, where the transmission networks are government owned, has there been concerns that insufficient international transmission capacity limits competition.[45]

7. Discussion and conclusions

In some industries, there are bottlenecks that make unregulated competition impossible or less effective than it would otherwise be. The bottlenecks can be infrastructure that is expensive to duplicate, but similar effects can also arise out of demand-side network effects: all customers want to belong to the network which "all" other customers are connected to. With a slight abuse of terminology, industries with significant bottlenecks will be "natural monopolies".

If one firm owns the bottleneck infrastructure, or controls access to a network with demand-side network effects, that firm will often hold considerable market power. Since unrestrained market power will result in inefficiencies, there is a motive for some type of regulatory intervention – even though regulation in itself will also result in inefficiencies.

A number of methods have been used in industries with bottleneck problems ("natural monopolies"), all with their respective pros and cons, including the following:[46]

- *Unregulated monopoly*. If competition from substitutes outside the markets are strong or if demand for other reasons is very elastic, if returns to scale are large and if regulation is likely to be costly, then it may be a reasonable alternative to let a natural monopoly to be unregulated.

- *Regulated monopoly*. If an unregulated monopoly is likely to result in an inefficient outcome (e.g., high prices), if regulation can be relatively efficient and if returns to scale are large – so that duplication is costly – then (consumer-price) regulation of a monopoly provider may be an alternative.

[44] During periods of rapid technological progress and innovation, private investments may be sufficient and even excessive. Helm argues that this is what we have seen in the telecom industry in recent years.
[45] Swedish Competition Authority, 2003.
[46] Other alternatives are horizontal separation and infrastructural clubs.

- *Government ownership*. A government-owned monopoly may be an alternative to a privately owned regulated monopoly, in particular if regulation is likely to be inefficient.

- *Franchise bidding*. If an unregulated monopoly is again likely to result in inefficiency (high prices) and if returns to scale are large, then franchise bidding may be an alternative to government ownership or a private monopoly under traditional regulation. In some situations, franchise bidding is informationally less demanding than regulation, since the price is set in a competitive bidding process, rather than by a regulator. An important disadvantage, however, is that the bidding for the franchise must be repeated regularly. The limited franchise tenure, on the other hand, may not provide good incentives for investments.

- *Vertical separation*. Sometimes one stage of production where the returns to scale are particularly large can be singled out. If this is the case, and if at the same time an unregulated monopoly would result in inefficiencies, vertically separation of that production stage from the others may be a good alternative. The infrastructural stage can either be government owned or privately owned and regulated. In fact, vertical separation may be useful to limit the negative consequences of regulation or government ownership.

- *Infrastructural access*. If one stage of production has large returns to scale, but if there are also large vertical synergies, then an alternative to vertical separation is to require the firm that controls that production stage to provide access to its rivals. The drawback is that regulation is likely to be more costly than under vertical separation.

The main objective of this paper is to explore the benefits and disadvantages of yet another approach: facilities-based competition (or competition in infrastructure). Since all of the above methods have drawbacks, it might seem attractive to try to come as close to the ideal competitive market and that, it would appear, means competition in infrastructure. Facilities-based competition has also been promulgated as the best long-term solution for the telecom industry. In fact, if returns to scale are relatively small in all production stages, at least compared to the costs of regulation, then it may indeed be a good policy to stimulate competition in all stages. However, if returns to scale are substantial, matters are more complicated.

When an industry such as the telecom industry is de-monopolised, competition will develop at different speeds in different segments. In most instances, the dynamic development of a market is best handled by the market itself. However, since the development in a bottleneck industry is dependent on the regulatory framework, the policymaker cannot completely sidestep the issue of where competition should first be introduced.

For at least two reasons, it seems natural to introduce competition in service provision before competition is introduced in the provision of infrastructure. Typically, returns to scale (or density) will be larger in the provision of infrastructure than in service provision. Consequently, duplication will be more expensive in infrastructure and, relative to its costs, the returns to introducing competition will be smaller. Furthermore, entry into services market will typically be associated with smaller sunk costs. This means that if one or several competitors enter the market, but competition turns out not to be viable, much less will be lost from an exit from services markets than from infrastructural markets.

The commonly held view that sector-specific regulation shall eventually be dismantled, the "sunset proposal", can be seen in light of the above argument. According to this view, competition will develop in successive stages. Initially, an industry may have been a regulated or state-owned monopoly. Then, in the first stage, entry will in principle be allowed, although facilities-based entry is unlikely to occur immediately. Instead, the entrants need to be assisted by the introduction of access regulation. In the second stage, the entrants will begin building their own infrastructure, but the market structure will still be asymmetric (one firm will be dominant). In this stage, there must still be some regulation, but it need not be as stringent as before. In the third stage, the market has become more symmetric and the industry is no longer dominated by a single firm. In this stage, sector-specific regulation can be dismantled.

Full-blown facilities-based competition, the third and final stage according to the above view, has the advantage that no regulation is needed and, consequently, the inevitable costs of regulation (such as regulatory capture, regulatory risk and bureaucracy costs; see Sections 3 and 4 and Bergman, 2002) can be avoided.

However, there are at least four problems with the above view. The first and most obvious problem is that when there are substantial returns to scale, facilities-based competition means wasteful duplication. Sometimes, the cost of duplication is worth incurring even in natural monopolies, because the benefits of competition throughout the whole "value chain" are so substantial. But sometimes economies of scale are large enough for them to be the primary concern, even if that means that one has to live with a less-than-perfect regulation of the monopoly bottleneck.

The second problem is that in industries where interconnection is essential, such as telecom and the payment-system industry, it is not necessarily true that regulation can be dismantled. In order to realise network benefits, subscribers must be able to make off-net calls and individuals and firms must be able to make payments to other banks and their customers. If the infrastructure is owned by two or more firms, this means that there must be two-way access. When the market is relatively symmetric, it is likely that two-way access will arise spontaneously. However, it is possible that a spontaneous two-way access regime will be anticompetitive.[47] Large firms may have incentives to foreclose – i.e., not interconnect with – smaller rivals and all firms may wish to design the mutual agreement so as to induce collusion (i.e., high final-customer prices). This is the so-called two-way access problem. (For an extensive treatment, see Armstrong, 2002, and Laffont and Tirole, 2000.) The general consensus appears to be that regulatory concerns will be smaller when ownership of the infrastructure is more symmetric (Valletti, 2003).

It should also be noted that interconnection for termination is more problematic than interconnection for origination. There is direct competition between firms that offer origination – i.e., phone services. The customer can chose the operator who offers the best prices and the best services. On the other hand, the active (calling and paying) party can typically not rely on competition when he or she wishes to make a call that requires interconnection for termination. The termination services must be provided by the operator which happens to be the one chosen by the receiving party, even if termination prices are excessive. Unless the receiving party cares for the welfare of the calling party – or fears

[47] While under one-way access, the main concern is one of foreclosure, now the main concern will be one of collusion. This problem is likely to be more pronounced when there is no competition for customers, as in international interconnection, and less serious when there *is* competition for customers, e.g., in domestic mobile-to-mobile interconnection.

receiving few calls- there is no incentive for the receiving party to be concerned with the termination fees. This asymmetry between interconnection for termination and interconnection for origination is reflected in the E-com directive, where the obligations laid on an operator are independent of the operator's market share in the termination market, but proportionate to the origination market shares. (In this context, it is worth recalling the trade-off between easy access and intense short-run competition, on the one hand, and less favourable access conditions and better incentives for long-run competition, on the other.)

The third problem with the "sunset proposal" is that the process towards balanced facilities-based competition is not an automatic one.[48] Clearly, the choice of access regimes in the first two stages will influence the development of competitive infrastructure. If the (one-way) access regime is "stringent" (favourable for the entrants), they will have little incentives to build their own infrastructure. If, on the other hand, the access regime is not stringent, there may be no entry at all. Possibly, a well-balanced regulation will result in a situation where facilities-based competition develops over time, but it is likely that for this to happen, the regulator must make active decisions. One example would be access prices that rise over time. Then, it would initially be advantageous to enter the services market. With time, however, it will become more and more advantageous also to invest in infrastructure (and less and less advantageous to be active in the services markets only). A side-effect of this type of dynamic regulation is that, as discussed above, for a period of time the incumbent's network business will also be more profitable – and possibly very profitable – before competition in infrastructure drives down margins.

Other examples of policies that actively promote facilities-based competition are LLUB and the format for granting mobile telephony licenses. It will be easier for competitive providers of infrastructure to enter if they do not have to provide the whole infrastructure; with LLUB they can compete by adding complementary features to the existing infrastructure. Alternatively, the entrants can compete in newly build areas or in new technology, such as new generations of mobile telephony.

Obviously, this kind of active policies requires that the policy maker is informed on where duplication is desirable, which in turn requires explicit or implicit estimates of scale economies, of the benefits of competition and of regulatory costs. A policy that favours competition in infrastructure may result in inefficient bypass or, if there is no or little entry, high prices due to high access prices. Conversely, a policy that favours competition in services may result in too little facilities-based entry and too little investment by the incumbent.

The fourth problem with the "sunset proposal" is that free-entry facilities-based competition may conflict with concerns for universal service. The new entrants will focus on low-cost high-demand customers, such as densely populated areas and big commercial customers. The incumbent will then be left with high-cost low-demand customers, with an associated access deficit. As discussed in Hultkrantz' chapter in this volume, there are methods to address the USO problem. However, since all regulation will inevitably result in some distortions of incentives, there will be costs associated with these methods too.

[48] Woroch (2002, p. 643) writes that "no inexorable, inherent tendency toward monopoly or toward competition can be discerned from the history of [the US telecom] industry. The past century witnessed several major transformations, first from unregulated monopoly to fierce competition, and then to regulated monopoly, and most recently to (de)regulated competition. Regulation and technological change played key roles in each case – in addition to luck and serendipity."

In the end, there is no regulatory panacea for natural monopolies. Regulation will always lead to inefficiencies, but the absence of regulation will also result in inefficiencies. There is never a perfect policy, only a "least bad" one, and even that may be elusive. At first glance, facilities-based competition may appear to be the Columbus' egg of natural monopolies. A closer look, however, reveals that this method is also imperfect. In particular, duplication of investments may not be sustainable in the absence of subsidies or other distortive incentives. On the other hand, sometimes the cost of duplication is worth incurring, because it makes competition more intense and because it reduces the regulatory burden.

References

Andersson, Peter, 2004, Tio år efter postmarknadens avreglering: effekter och reformförslag, rapport för Konkurrensverket.

Armstrong, 2002, The Theory of Access Pricing and Interconnection, in Martin E. Cave, Sumit K. Majumdar and Ingo Vogelsang (Eds.) *Handbook of Telecommunications Economics*, Vol. 1, North-Holland, Amsterdam.

Armstrong, Mark 2004, *Competition in Two-Sided Markets*, mimeo, University College London.

Armstrong, Mark, Simon Cowan and John Vickers, 1994, *Regulatory Reform. Economic Analysis and British Experience*, MIT Press, Cambridge, USA

Averch, Harvey and Leland Johnson, 1962, Behavior of the Firm under Regulatory Constraints, *American Economic Review,* 52, 1052-1069.

Bergman, Lars, Chris Doyle, Jordi Gual, Lars Hultkrantz, Damien Neven, Lars-Henrik Röller and Leonard Waverman, 1998, *Europe's Network Industries: Conflicting Priorities Telecommunications Monitoring European Deregulation*, CEPR, London. (A Swedish translation was published in 1999 by SNS Förlag, Stockholm, under the title *Europas nätverksindustrier. Telekommunikationer. Avregleringen i Europa*.)

Bergman, Mats, 2002, Lärobok för regelnissar. En ESO-rapport om regelhantering vid avreglering, Ds 2002:21, Stockholm.

Bergman, Mats, 2004, *When Should an Incumbent Be Required to Provide Access Under the General Competition Rules?*, Stockholm University, mimeo.

Carlton, Dennis W. and Jeffrey M. Perloff, 2004, Modern Industrial Organization, Pearson/Addison Wesley, Boston.

Cave, Martin and IngoVogelsang, 2003, How Access Pricing and Entry Interact, *Telecommunications Policy*, 27, 717-727.

Connor, John M, 2003, *International Price Fixing: Resurgence and Deterrence*, Purdue University, IN, mimeo.

Falch, Morten, 2001, Cost and Demand Characteristics of Telecom Networks, in *Telecom Reform: Principles, Policies and Regulatory Practices*, Ed. William H. Melody, available at http://www.lirne.net/resources/books/books.htm.

Fuss, Melvyn A. and Leonard Waverman, 2002, Econometric Cost Functions, in Martin E. Cave, Sumit K. Majumdar and Ingo Vogelsang (Eds.) *Handbook of Telecommunications Economics*, Vol. 1, North-Holland, Amsterdam.

Garzantiti, Laurent and Francesco Liberatore, 2004, Recent Developments in the European Commission's Practice in the Communications Sector, *European Competition Law Review*, 25, 234-240.

Gonec, Rauf and Giuseppe Nicoletti, 2000, *Regulation, Market Structure and Performance in Air Passenger Transportation,* OECD, Economics Department Working Paper No. 254.

Guibourg, Gabriela, 2001, *Interoperability and Network Externalities in Electronic Payments*, Sveriges Riksbank Working Paper Series, No.126, Stockholm.

Helm, Dieter, 2001, The Assessment: European Networks – Competition, Interconnection, and Regulation, *Oxford Review of Economic Policy*, 17, 297-312.

Joskow, Paul L., 1987, Contract Duration and Relationship-Specific Investments: Empirical Evidence from Coal Markets, *American Economic Review*, 77, 168-85.

Katz, Michael and Carl Shapiro, 1985, Network Externalities, Competition, and Compatability, *American Economic Review,* 75, 424-440.

Laffont, Jean-Jacques and Jean Tirole, 1993, *A Theory of Incentives in Procurement and Regulation,* MIT Press, Cambridge, USA.

Laffont, Jean-Jacques and Jean Tirole, 2000, *Competition in Telecommunications,* MIT Press, Cambridge, USA.

Lerner, Josh and Jean Tirole, 2004, Efficient Patent Pools, *American Economic Review*, 94, 691-711.

Liu, Zhiqiang, 2001, Efficiency and Firm Ownership: Some New Evidence, *Review of Industrial Organization,* 19, 483-498.

Milgrom, Paul and John Roberts, 1993, *Economics, Organization and Management*, Prentice Hall, Englewood Cliffs, NJ.

Noam, Eli M., 1985, Economies of Scale and Regulation in CATV, in Michael A. Crew (Ed.), *Analyzing the Impact of Regulatory Change in Public Utilities,* Lexington Books, Lexington, MA.

Nordberg, Mikael, 2004, mimeo

OECD, 2001, Restructuring Public Utilities for Competition, Paris.

Oldale, and Jorge Padilla, forthcoming, Jacob's Ladder…. , in the Swedish Competition Authority's anthology *The Pros and Cons of Antitrust in Deregulated Markets*.

Owen, Bruce M. and Peter R. Greenhalgh, 1986, Competitive Considerations in Cable Television Franchising, *Contemporary Policy Issues*, 4, 69-79.

Posner, Richard A, 2003, *Antitrust Law: An Economic Perspective*, University of Chicago Press, Chicago.

PTS, 2003, *Svensk telemarknad första halvåret 2003*, report PTS-ER-2003, 18 December 2003.

Sung, Nakil and Michael Gort, 2000, Economies of Scale and Natural Monopoly in the U.S. Local Telephone Industry, *Review of Economics and Statistics*, 82, 694-97.

Swedish Competition Authority, 2003, *A Powerful Competition Policy*, report 2003:1 from the Nordic Competition Authorities.

Valletti, Tommaso M., 2003, The Theory of Access Pricing and its Linkage with Investment Incentives, in *Telecommunications Policy*, 27, 659-675.

Viscusi, Kip W., John M. Vernon and Joseph E. Harrington, Jr, 2000, *Economics of Regulation and Antitrust*, MIT Press, Cambridge, MA.
Webb, Kent G., 1983, *The economics of cable television*, Lexington Books, Lexington, MA

Winston, Clifford, 1998, U.S. Industry Adjustment to Economic Deregulation, *Journal of Economic Perspectives*, 12, 89-110.

Woroch, Glenn A., 2002, Local Network Competition, in Martin E. Cave, Sumit K. Majumdar and Ingo Vogelsang (Eds.) *Handbook of Telecommunications Economics*, Vol. 1, North-Holland, Amsterdam.