In contrast, the Chinese policy approach is more influenced from the macro level and likely to be nonincremental. In particular, China has early on recognized the importance of telecommunications for economic growth and has therefore pushed technological and market advances. The remarkable property of the Chinese approach is the parallel existence of several telecommunications carriers owned by the central state. How these companies with common ownership compete with each other is certainly worth an academic investigation. Xia (2017) points out that China has specifically promoted competition, while containing private participation in network operations and that it has been able to separate ownership from regulatory functions in government. Liu and Jayakar (2012) do, however, emphasize the paradoxical regulator-owner interface. In contrast to the network infrastructure provision, service providers such as mobile virtual network operators (MVNOs) and telecommunications equipment manufacturing are allowed to be in private hands.

1.4 Overview

The next section develops specific economic concepts associated with network industries. It is followed in Section 3 by regulatory approaches based on monopoly. These sections concentrate on monopoly, in spite of the fact that competition today is present in all network industries. However, monopolistic bottlenecks persist in core areas. The economic and regulatory treatment of these core areas is more complex and builds on insights from the simple monopoly approach, which therefore comes first. Section 4 analyzes those competitive developments and their regulatory treatment. Section 5 addresses some special issues of telecommunications. Section 6 deals with the current and upcoming issue of deregulation. Section 7 concludes.

2 Economic Concepts Associated with Network Industries

Because of the specific economic features of network industries, a number of economic concepts have been developed for their study. Although these concepts have general applicability throughout the economy, they were developed here first and have found their widest application in network industries. These concepts refer to costs and demands. This section also includes the resulting welfare concepts for a normative analysis.

2.1 Single-Product Cost Concepts

The first major cost concept concerns economies of scale, which define the cost advantage of large networks over small networks and lead to natural monopolies in a single-product setting. Second, there are various concepts associated with

networks as multiproduct firms. These concepts include incremental costs and stand-alone costs, which are necessary for defining economies of scope and cross-subsidies. Together with economies of scale, economies of scope lead to natural monopolies in a multiproduct setting. The concept of average cost, which helps define economies of scale in the single-product case, is no longer well defined for multiproduct firms and is therefore replaced by ray-average costs.

Under a single-product firm, economies of scale mean per unit cost advantages from producing more of the same product; that is, average cost declines as the output increases,

$$\frac{dAC(Q)}{dQ} < 0.$$

Here 'Q' stands for the quantity of output and 'AC' for average cost. Also, under economies of scale, the elasticity of cost w.r.t. output, σ_C, is less than 1,

$$\frac{MC}{AC} = \sigma_c = \frac{dC(Q)}{C(Q)} \Big/ \frac{dQ}{Q} < 1.$$

Here C(Q) is the cost function and MC stands for marginal cost. If $\sigma_c = 1$, there are constant costs or constant returns of scale. If the inequality is reversed, there are diseconomies of scale.

Where do scale economies come from? It is easy to envisage a constant cost industry, where a doubling of all inputs leads to doubling of output. However, both economies of scale and diseconomies of scale are harder to explain. There are four common explanations for economies of scale. First, some inputs come in lumps. Such indivisible inputs lead to downward-sloping average cost curves over some range, until the input reaches its capacity. Then, as output increases, another indivisible input has to be added, leading to a jump in average cost and then again to declines. As output increases further, this leads to average cost ratcheting with declining peaks. A second explanation for economies of scale is the 2/3 rule for the relationship between surface and volume of containers. This holds, for example, for ducts that carry fibre-optic cables. Here the 2/3 rule would apply to the size of ducts, while lumpiness and sunk costs hold for laying the ducts in the ground. The third and most common advantage is the division of labor made famous by Adam Smith. A fourth explanation concerns quantity rebates on input prices. This also alerts to the fact that economies of scale and returns to scale are related but not the same concepts. Economies of scale are a cost concept, while returns to scale are a production function concept. This explanation naturally begs the question where these quantity rebates come from. Here again economies of scale can be a major reason, while buying power could be another.

What are the specific reasons for economies of scale in network industries? First, networks are composed of links and nodes that tend to be capital goods with lumpy characteristics. Second, networks have to either link subscribers to a source or several sources or to each other. Switched nodes then allow for savings on links so that the total number of links can be much smaller than if every subscriber were directly linked to the source or to each other. These savings increase dramatically in a factorial way with the number of subscribers.

For networks, a related concept to economies of scale are economies of density. Such economies relate to the fact that for a given number of subscribers the cost of a network with smaller geographic coverage will have lower cost. Thus, a telephone network in a densely populated city will have lower cost per subscriber than a network in a large rural area with the same number of subscribers. The network links in the city will simply be shorter (although this could be compensated for by higher real estate prices and wages in the city).

Although economies of scale and sunk costs are in principle independent of each other, economies of scale in network industries are commonly associated with sunk costs, such as those incurred by digging up the ground for installing ducts or lines. Sunk costs are defined by the property that the costs of an input, once they have been spent, cannot be recovered other than by using the input for the particular dedicated output. In other words, there is no functioning second-hand market for the particular input. The sunk cost property increases the risk and thereby the cost of investment and can create a barrier to entry.

2.2 Single-Product Natural Monopoly Concepts

Closely related but not identical to economies of scale is the natural monopoly property. In the traditional view, it attempts to answer the question of what the cost-minimizing market structure is. This *supply-side natural monopoly* (= *classic natural monopoly*) means that total costs of industry output is less when produced by a single firm than by any number 'N' of firms greater than one. In other words, it's cheaper to produce all the outputs in a single firm than in more than one firm. A firm represents a *natural monopoly* if its cost function is *sub-additive* over all relevant outputs,

$$C\left(\sum_{i=1}^{N} Q_i\right) < \sum_{i=1}^{N} C(Q_i), N \geq 2.$$

The classic natural monopoly is clearly caused by cost advantages of being large. However, natural monopoly can still exist even if there are diseconomies of scale (or scope) over some range of output(s). If this range is sufficiently

small, a single firm will have lower cost of the total market output than two or more firms, each of which does do not exhaust its scale economies.

While the classic natural monopoly is described in all textbooks on public utility regulation, a newer demand-related natural monopoly concept is rarely mentioned, although it is of potentially major importance for modern network industries. Direct and indirect network effects have been called *demand-side economies of scale* (Shapiro & Varian, 1999). They can give rise to a *demand-side natural monopoly*. Such a natural monopoly characterizes the consumer-surplus-maximizing market structure at a given price. It is associated with a *super-additive* demand function (versus the *sub-additive* cost function for supply-side determinants of natural monopoly). It is related to endogenous sunk costs, which are global and grow with market size. Super-additive demand is relevant for Internet-related industries, such as social networks.

The demand-side natural monopoly was first defined by Shaffer (1983). Begin by defining the inverse demand function $P(Q)$ to be strictly super-additive if and only if for all firms i and for $Q = \sum_{i=1}^{N} Q_i$:

$$P(Q) > \sum_{i=1}^{N} P(Q_i), Q_i > 0, N \geq 2 .$$

This condition says that consumers are willing to pay a higher price for an amount Q of a commodity if it is produced by a single firm than if it is produced by any combination of two or more firms. From this follows that a single-product industry with constant returns to scale is a strict (demand-side) natural monopoly if and only if its inverse demand function is strictly super-additive. A sufficient condition for super-additivity is that average revenue (or inverse demand) increases in scale (Shapiro & Varian, 1999). This can happen if positive network effects from an increased subscribership outweigh the price effects of an otherwise downward-sloping demand.

The downward-sloping thin inverse demand curves $P(Q,Q_i)$ in Figure 1 represent market demands as a function of price for a given expected number of subscribers, Q_i. The demands shift outward as i increases. The expectations, however, are only fulfilled where $Q_i = Q$. Linking these points of fulfilled demands yields the fulfilled expectations demand, which over some range is upward-sloping.[4] In this upward-sloping part, the network effect of increasing demand is stronger than the conventional law of decreasing demand. As more subscribers have joined the network, the network effect decreases, leading to a downward-sloping segment of the fulfilled expectations demand.

[4] This insight originally goes back to Rohlfs (1974). See also Mitchell & Vogelsang (1991).

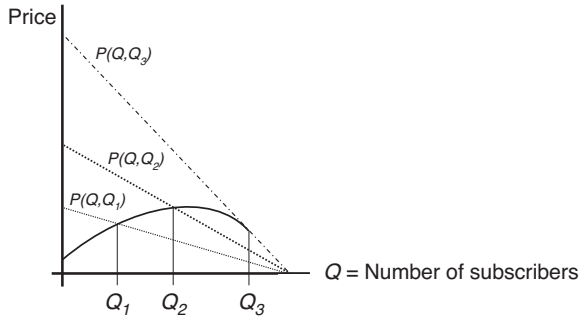

Figure 1 Fulfilled expectations demand

In principle, for many prices there exist two quantities that could be market outcomes. The equilibrium points in the upward-sloping portion of the fulfilled expectations demand are unstable, while those in the downward-sloping portion are stable. Unstable here means that if the actual demand turns out to be higher (lower) than expected, a new equilibrium will result with a higher (lower) number of demanded subscriptions than before, which means that the number of subscribers will again be underestimated (overestimated). Overall, market participants would have an interest in settling for points in the downward-sloping section of demand, which may require coordination in the form of nudging or some financial incentives.

Under demand-side natural monopoly (and with no diseconomies of scale in production) a single supplier would be the efficient market structure. However, in reality, consumers typically have heterogeneous tastes regarding networks. Thus, the demand-side economies of scale compete with the benefits of product differentiation if differentiated networks are incompatible with each other. Thus, consumers will weigh the benefits from joining a larger network with those from joining a network that is more to their taste. If the network externalities are stronger than the perceived benefits of product differentiation then there may still exist a demand-side natural monopoly. Furthermore, demand-side economies and supply-side economies may come together, thereby potentially creating strong natural monopoly conditions. Google's search engine may be a case in point. Even if network services are homogeneous, a demand-side natural monopoly does not require monopoly provision if interconnection between networks or multi-homing is feasible and cheap.

2.3 Relevant Concepts for Multiproduct Firms

Although microeconomics concentrates on single-product firms, the economy is actually dominated by multiproduct firms. In fact, it is hard to name any single-product firms. Even firms that seem to produce a single homogeneous product usually offer different varieties of it. In contrast, all the common enterprises around us are multiproduct firms. This holds, in particular, for services such as those offered by network industries. Why are multiproduct firms particularly common in service industries that seem to offer homogeneous products like electricity? Services are typically already differentiated by time and location. Because it cannot be stored, electricity sold during the day is no close substitute for electricity sold during the night. Likewise, a telephone call between cities A and B is no close substitute for a call between cities C and D.

Multiproduct firms require a special approach to profit maximization and welfare maximization, because different products have costs in common and because their demands interact. For example, the average-cost concept developed for single-product firms does not work for multiproduct firms. Why? It is because average cost cannot be defined fully generally, since output now is a vector and one cannot divide total cost by several different output quantities (or divide a scalar by a vector). How then do we define average costs for these firms? One needs a more restricted definition. The first and most relevant of these concepts is called *ray average cost* (RAC). Assume the total cost for a two-product firm is $C(Q_1, Q_2)$, meaning that the firm makes two products, *1* and *2* with the quantities q_1 and q_2. Further assume that the two outputs are produced in a constant ratio $\varsigma_1 : \varsigma_2,\ s.t.\ \varsigma_1 + \varsigma_2 = 1$. Then the set of outputs is defined implicitly from the equations $Q_1 = \varsigma_1 Q$ *and* $Q_2 = \varsigma_2 Q$. Ray average cost is now defined as:

$$RAC(Q) = \frac{C(\varsigma_1 Q, \varsigma_2 Q)}{Q}.$$

RAC assumes that outputs are produced in fixed proportions. The RAC can vary for every product ratio and product level. Thus, RAC captures average costs along a ray from the origin. Further on, we will define the simpler concept of average incremental cost.

As in the single-product case, economies of scale for multiple products relates to (ray) average cost. If we increase all outputs proportionally, say by a factor of v, then multiproduct economies of scale are defined by

$$\frac{\partial RAC(vQ)}{\partial v} < 0.$$

This means that we have decreasing ray average cost. In this case, multiplying by v makes the output proportions stay the same.

In contrast to average cost, the marginal cost for product i of a multiproduct firm is well-defined as

$$MC_i = \frac{\partial C(Q_1, Q_2)}{\partial Q_i}, \; i = 1, 2.$$

Very useful concepts for costing and pricing in network industries are *incremental cost* and *stand-alone cost*. Again, consider the two-product case with total cost: $C(Q_1, Q_2)$.

The *stand-alone cost* is then defined as the cost of only producing one product:

$$SAC(Q_1) = C(Q_1, 0), \; SAC(Q_2) = C(0, Q_2).$$

In contrast, the incremental cost is the cost of adding a product if the firm is already producing one product.

Thus, the incremental cost of product *1* is the total cost minus stand-alone cost of the other product *2*.

$$IC(Q_1) = C(Q_1, Q_2) - C(0, Q_2).$$

If a producer starts producing only product *2*, then the incremental cost of product *1* can be interpreted as the extra cost the producer has to incur if he starts to produce both products.

The *average incremental cost* of product *1* now is defined as:

$$AIC(q_1) = \frac{C(Q_1, Q_2) - \; C(0, Q_2)}{Q_1}.$$

Declining AIC defines *product-specific* economies of scale:

$$\frac{\partial AIC(Q_1)}{\partial Q_1} < 0.$$

A very useful concept characterizing the economies achieved by having multi-product rather than single-product firms is that of *economies of scope*. Economies of scope mean that it is cheaper to produce a number of different products together than separately. They are also called synergies.[5]

In the two-product case, *economies of scope* exist if the sum of stand-alone costs is greater than total cost:

[5] Such synergies can also induce separate firms to share assets or to coinvest rather than to merge fully.

surplus in the multiproduct case is easily defined for the case of products that are independent in demands. In this case the multiproduct consumer surplus is simply the sum of the single-product consumer surpluses. Thus, in the multi-product case, if products are independent from each other (cross elasticity is zero), then

$$V(\boldsymbol{P}) = \sum_{i=1}^{N} CS(P_i).$$

In general this, however, no longer holds if the products are substitutes and/or complements. If the cross elasticities do not equal zero, we have to consider the effect of a price change of P_i on $CS(P_j),\ i \neq j$. In that case the change of the price of one product shifts the demand for the other product(s). Thus, these shifts generate further consumer surplus additions or reductions that have to be taken into consideration. Since these additional changes in general depend on the order in which price changes are done, the multiproduct consumer surplus very often is no longer unique (i.e., it is path dependent). It is only unique if the cross-derivatives of demand for goods i and j are the same as between goods j and i. This holds if there are no income effects.

2.5 Welfare Benchmarks for Policies

For establishing a simple welfare benchmark in monopoly, we assume a static, single product, and a full information environment.

As indicated previously, the regulator maximizes *social surplus* with respect to price, which results in price equaling marginal cost:

$$\max_{P} W(P) = \pi(P) + CS(P) = PQ - C(Q) + \int_{\tilde{p}}^{\infty} Q(P)dP$$

$$F.O.C.\ w.r.t.\ P\text{:}\ \ Q + P\frac{\partial Q}{\partial P} - \frac{\partial C(Q)}{\partial Q}\frac{\partial Q}{\partial P} - Q = 0$$

$$\Rightarrow\ \left(P - \frac{\partial C(Q)}{\partial Q}\right)\frac{\partial Q}{\partial P} = 0$$

$$\Rightarrow\ P = MC.$$

However, due to economies of scale, $P = MC \rightarrow P < AC$. Thus, in order to achieve the optimal price, the regulator would have to *subsidize* the firm to make up for the loss. However, several problems may arise from using a subsidy in order to achieve efficient pricing. First, a subsidy may give the firm wrong incentives such that the firm has little motivation to lower its cost. Second, the

total consumer willingness to pay may be less than the cost of production. Third, and particularly important, the subsidy comes from the government, which raises money through taxes, through profits of state-owned firms, by issuing debt, or through inflation (by printing money). In all these cases, raising the money for the subsidy could create major distortions in other markets thereby causing welfare losses. These losses can be quite high. Thus, by using subsidies, the government is improving the efficiency in one market while creating a potentially much larger inefficiency in another market. Fourth, the availability of subsidies may induce bribes to get them.

Instead, as a result of all these factors, it is preferable for the regulator to maximize net social surplus under a *break-even constraint*. In the single-product case, with economies of scale, this will simply lead to average cost pricing.

In the multiproduct case, where average costs are not well-defined, the policy rule concerning the welfare-maximizing monopoly prices, subject to a constraint on the monopoly's profit being nonnegative, is called *Ramsey pricing*. It is designed to maximize social welfare with the least distortions across markets.

The setup for Ramsey pricing is as follows:

Total cost: $C(Q_1, Q_2, \ldots, Q_N)$.

Demand for each product: $Q_i(P_1, P_2, \ldots, P_N), i = 1, 2, \ldots N$.

We differentiate between the cases of independent and interdependent demands.

Under independent demands, $Q_i(P_1, P_2, \ldots, P_N) = Q_i(P_i)$, for all $i = 1, 2, \ldots N$.

In this case, the regulator's problem becomes

$$\max_{(P_1, P_2, \ldots P_N)} (1+\mu)\left[\sum_{i=1}^{N} P_i \cdot Q_i - C(Q_1(P_1), Q_2(P_2), \ldots Q_N(P_N))\right] + \sum_{i=1}^{N} CS_i(P_i)$$

$$s.t.\ \pi \geq 0$$

$$\Rightarrow \quad \text{Lerner Indices } LI_i \equiv \frac{P_i - \frac{\partial C}{\partial Q_i}}{P_i} = -\frac{\mu}{1+\mu} \cdot \frac{1}{\varepsilon_i}$$

$$\Rightarrow \quad \frac{LI_i}{LI_j} = \frac{\varepsilon_j}{\varepsilon_i}.$$

Thus, in the case of a nonbinding break-even constraint with $\mu = 0$, we get marginal cost pricing. The constraint will be nonbinding if economies of scale are exhausted. The Lagrange multiplier will become infinite if the unconstrained profit-maximizing monopoly will just be able to break even. In general, under Ramsey pricing with independent demands and a binding constraint, the

ratio of markups of any two products equals the inverse ratio of their elasticities. If a product has the relatively higher (lower) absolute value of the elasticity, then its price markup should be lower (higher).

In contrast, under interdependent demand, we have $Q_i = Q_i(P_1, P_2, \ldots P_N)$, $i = 1, 2, \ldots N$.

Note here that one cannot add consumer surpluses separately. Thus, in this case, the regulator's problem becomes

$$\max_{(P_1,\, P_2,\, \ldots P_N)} \mathcal{L} = (1+\mu)\left[\sum_{i=1}^{N} P_i \cdot Q_i - C(Q_1, Q_2, \ldots Q_N)\right] + V(P_1,\ P_2, \ldots P_N)$$

$$s.t.\ \pi \geq 0$$

$$F.O.C.\ w.r.t.\ P_i:$$

$$\frac{\partial \mathcal{L}}{\partial P_i} = (1+\mu)\left[Q_i(P_1, \ldots P_N) + \sum_{j=1}^{N} P_j \cdot \frac{\partial Q_j}{\partial P_i} - \sum_{j=1}^{N} \frac{\partial C}{\partial Q_j} \cdot \frac{\partial Q_j}{\partial P_i}\right]$$

$$- Q_i(P_1, \ldots P_N) = 0$$

$$\Rightarrow \textit{Lerner Indices } L_i \text{ with} \quad LI_i \equiv \frac{P_i - \frac{\partial C}{\partial Q_i}}{P_i} = -\frac{\mu}{1+\mu} \cdot \frac{1}{\eta_i}.$$

Here η_i is the *super elasticity* of product i. Super elasticities are combinations of direct and cross elasticities that capture the direct and indirect effects of price changes. The super elasticities for a two-product case are as follows. Denote $\varepsilon_k = \frac{\partial Q_k}{\partial P_k} \cdot \frac{P_k}{Q_k}$ = ordinary elasticity and $\varepsilon_{kl} = \frac{\partial Q_k}{\partial P_l} \cdot \frac{P_l}{Q_k}$ = cross elasticity.

Then

$$\eta_1 = \varepsilon_1 \frac{\varepsilon_1\varepsilon_2 - \varepsilon_{21}\varepsilon_{12}}{\varepsilon_1\varepsilon_2 + \varepsilon_1\varepsilon_{12}}$$

$$\eta_2 = \varepsilon_2 \frac{\varepsilon_1\varepsilon_2 - \varepsilon_{21}\varepsilon_{12}}{\varepsilon_1\varepsilon_2 + \varepsilon_2\varepsilon_{21}}.$$

The sign of η_i can be either positive or negative. To see this, consider the two-product case: If the two products are *substitutes* and we increase P_1 then, due to the substitution effect, Q_2 increases. This usually leads to a higher optimal Ramsey markup for that particular good. If the two products are *complements*, increasing P_1 lowers Q_2. This usually leads to a lower optimal Ramsey markup for that particular good. The existence of complements is a necessary, but not sufficient, condition for super elasticities to be negative. Negative super elasticities would lead to negative Ramsey markups for those services. This possibility is closely related to the well-known similar result in two-sided markets.