1 Introduction

1.1 What Are Network Industries and What Makes Them Interesting Topics in Economics?

The economics and regulation of network industries is an old-fashioned topic that has been totally modernized. What used to be "public utilities" is now called "network industries." This name change signifies both a change in the scope of the relevant industries and a change in their character and of the inquiry into their properties. A typical public utility owned an infrastructure that was made available to end users via regulated monopoly markets. The regulation was justified by a public interest in the services and the required containment of monopoly power. However, regulation was rarely efficient. The resulting critique of traditional regulation naturally induced economists both to recommend deregulation and to look for improvements of regulation.

In a network industry, each company owns a network that is defined by links and nodes. The network connects users with each other. The users can be of the same types, such as in the case of telephone subscribers, or of different types, such as electricity generators and loads or Internet content providers and end users. In the second case, the network acts as a platform that deals with two-sided or multi-sided markets of users. In the Internet age, more and more networks have become interactive rather than being simple distribution networks.

Network industries have several common features that make them challenging and exciting topics of theoretical and empirical inquiry. The combination of the following features means that network industries are subject to various policy interventions.

First, network industries are of vital importance to the economy as infrastructures and "general purpose technologies" (Bresnahan & Trajtenberg, 1995) that the whole population needs and that spill over from individual users to the whole economy. This, for example, justifies universal service policies making the network services available to everyone.

Second, as recognized early on, networks exhibit sunk costs and supply-side economies of scale and scope, which favor large firms supplying many services together. Network industries have therefore traditionally been associated with market power.

Third, a newer subject of inquiry and possibly the reason for the name change from public utility to network industry is the presence of network effects, which are demand-side economies of scale and scope. Without either supply-side or demand-side economies there would be no reason for

firms to have networks. They would then link up with each customer individually. Similar to supply-side economies, demand-side economies can lead to market power, but they can also be associated with other allocative distortions. The latter could be the result of direct network effects, which are externalities. For example, a telephone network is worth little if it has only a single subscriber. It gains its value from the presence of many subscribers who can talk to each other. Thus, the decision of a person to become a subscriber benefits others who are not part of that decision. Direct network effects can also be negative, such as in the case of congestion in transport networks, where an additional user increases congestion.

In recent years, indirect network effects have gained increasing importance, largely due to the Internet. Indirect network effects arise from decisions of one type of network user that affect other types of network users. For example, the addition of new newspaper subscribers has a positive effect on advertisers. In the past, due to many subscribers, newspaper subscriptions were cheap, because advertisers largely paid for them. Today they are expensive, because advertisers pay less due to the diminished subscriber base. This is an example of two-sided markets, where the network acts as a platform. The economics of two-sided and multi-sided platforms therefore has a lot in common with the economics of network industries. Since Belleflamme and Peitz (2015) and Comino and Manenti (2014) have skillfully surveyed platforms, we only present some of the main results relevant to network industries and their regulation.

Do all network industries exhibit network effects? Examples outside the telecommunications sector include congestion for transport networks and Kirchhoff's laws for electricity transmission. The expansion of a network may create network effects even in the absence of an increased number of users, although those may join following the expansion. For example, the build-out of a road or railroad system opens up demand opportunities that will attract more users and will usually, but not always, reduce congestion.[1] "If you build it, they will come." These network effects and Kirchhoff's law show the interaction between the supply and the demand side.

While positive network effects favor large network providers, they may also be realized through interconnection between standardized networks. Thus, interconnection and standardization play an important role for network industries and the policies associated with them.

[1] The possibility that adding new capacity can increase congestion in a transport network is known as "Braess's Paradox" (Braess, 1969).

A fourth important property of networks is that they generally operate in service industries, which means that their services are consumed at the same time they are produced. As a result, services generally cannot be stored.[2] Since networks are capital goods with capacity limits that depend on the amount invested in them, high-capacity utilization is essential for their profitability. This can, among others, be accomplished through time-dependent pricing techniques and technology choices that reflect the time duration a particular capacity is being used. Furthermore, price discrimination and sophisticated pricing techniques are more appropriate and more easily used in service industries than otherwise.

1.2 The Relationship between Networks and the Services Provided over Them

Networks can be used for the provision of various (goods or) services, such as telephony, cable TV and data services for modern (tele-)communications networks. The networks themselves are the infrastructures over which services are delivered. The networks often require public resources, such as airwaves or rights of way, and have to be managed either by themselves or by public agencies, such as air traffic controllers.

Because of the complementarity of the networks and the services provided over them, the issue of vertical integration vs. separation has arisen. When networks as particular industries first evolved, they were usually vertically integrated in the sense that they produced the inputs necessary for the service provision, and transported the services over the network and sold them to end users. For electricity network providers, this meant that they generated electricity, transported it over the electricity network and sold it at retail. Relative to the network as the core activity, this meant that generation as the upstream activity was integrated, as was retail as the downstream activity. Over time, in electricity, two separate core network activities emerged in the form of transmission and distribution networks, which transport electricity at different voltage levels and over different distances. Similarly, telecommunications networks included long-distance and local networks with different economic properties. Railroads used to be vertically integrated by including the rail tracks, stations and the trains and equipment.

Vertical integration can possess efficiency advantages that make such institutional arrangements more efficient than vertical separation. However, since

[2] Note that goods, such as natural gas, which flow over networks can be stored, while the transportation service of the pipeline system cannot. The storage of such goods can, however, substitute for the infeasible storage of the transport service.

core networks mostly exhibit strong economies of scale and scope, while the upstream and downstream activities often do not, competition may be possible for the noncore activities. The advantages of competition on these production stages may then have to be traded off against the disadvantages of vertical separation associated with such competition for the noncore activities.

There are two additional complications in the new developments in network industries. The first is that some of the core activities have exhausted or no longer exhibit economies of scale and therefore can be duplicated without too much loss of production efficiencies. The more this is true, the less there is a policy case for vertical separation (for given vertical economies). The second is that noncore activities, such as "over the top" (OTT) services by nonnetwork owners that are provided over the public Internet, can (a) gain bottleneck properties and/or (b) can compete against activities of network providers.

Based on the strong presence of economies of scale and scope and of network effects, the traditional provision of network industries has been by monopolies with vertical integration of network infrastructure and services. Networks in many countries were state-owned, such as telephone, electricity, gas, water and railroad networks. Today, network industries are characterized by general competition for services and some competition for network infrastructure. Vertically integrated firms compete alongside vertically separated firms, which often depend on essential network inputs from the integrated networks.

Current hot topics in network industries include the emergence of smart electricity distribution grids, electricity storage, net neutrality for Internet service provision of content, the sharing economy, over-the-top services (OTT) in (tele-)communications networks, and the emergence of 5G in mobile communications. We put special emphasis on the telecommunications sector, because it has all the network features and is the most researched network industry. This Element focuses on some models and applications and does not extensively cover the empirical and policy literature.

1.3 Relevant Economic Policies

Network industries have been subject to many policy interventions. The current Element takes a normative approach to such policies but recognizes that there can be conflicts between the normative and the positive approach, which can be relevant for policy recommendations (Briglauer et al., 2019). Since we concentrate on specific industries, only microeconomic policies are of concern.

The main policy issues for network industries concern market power and externalities/network effects. Most countries use two types of policies for dealing with market power. They are competition policy (antitrust), as the

general policy relevant for all industries, and economic (industry-specific) regulation, which specializes on particular industries. There exist also two types of policies dealing with externalities: social regulation as the general policy relevant for all industries and industry-specific regulation that sometimes can be combined with economic regulation. Examples of industry-specific policies dealing both with market power and network effects are interconnection regulation for telephony, net neutrality regulation of Internet service providers (ISPs); and radio spectrum regulation for mobile networks, TV and other services.

Competition policy governs all industries, trying to preserve and enhance competition in markets, usually for the long-term benefits of consumers. Competition is threatened by market power, collusion and fraudulent behavior. The relevance of competition policy to network industries is mostly concentrated on issues of monopolization, foreclosure, predation, tying/bundling and mergers. Collusion is currently not a big issue for network industries but that may change. In particular, coinvestment and asset sharing to lower costs are currently hot topics in telecommunications, and they can be associated with collusion between the partners of such undertakings (Krämer & Vogelsang, 2016). The importance of competition policy increases with the increase in competition in network industries and with the increase in the complexity of network industry structures. There is no policy void if industry-specific regulation is abandoned. In particular, competition policy is the fallback in case of deregulation. This happens with the advance of competition in telecommunications and to some degree in electricity. Interestingly, through these developments, competition policy is becoming more "regulatory" (Geradin & Sidak, 2005). Sometimes overlaps or conflicts occur between competition policy and regulation. Most relevant in this regard has been the 2004 US Supreme Court case *Verizon v. Trinko*, where Court ruled against allowing an overlap.[3]

Since industry-specific regulation of private enterprises in its modern form was first practiced in the USA, and since other countries have adapted to it, we put particular emphasis on the US model and briefly touch upon the European Union (EU), India and China.

US regulation is performed by "independent" commissions. Independent here means that the regulatory agency is somewhat independent of the government and has some executive, legislative and judiciary powers. The regulatory agency is headed by commissioners, who are supported by a sometimes-large staff. Commissioners can be appointed by the head of government or elected by popular vote. There are federal and state commissions for the regulated network

[3] *Verizon Communications v. Law Offices of Curtis V. Trinko*, LLP, *540 U.S. 398* (2004).

industries. Federal commissioners are appointed by the US President and confirmed by the Senate, while state commissioners can either be appointed by the Governor or be elected. The number of staff members varies between 3 for small state commissions and 2,500 for large federal commissions, such as the Federal Energy Regulatory Commission (FERC) or the Federal Communications Commission (FCC).

The division of labor between federal and state commissions is guided by the Commerce Clause and the Supremacy Clause of the US Constitution. The Commerce Clause makes sure that activities that directly concern several states (inter-state commerce) are dealt with at the federal level, while activities concerning a single state (intra-state commerce) are dealt with at that state level. While this sounds like a clear rule, it can be quite fuzzy for network industries that are interconnected across states.

The Supremacy Clause allows Congress to pass laws that concern the whole USA, even if each activity is restricted to single states. This has, for example, occurred when the Telecommunications Act of 1996 regulated wholesale access to local telephone lines. This was done in order to have a unified approach across the country.

The regulatory activities can be classified into adjudication and rulemaking. Under adjudication, the commissioners will decide about a single case, which can be an application by the regulated firm for a price increase or the complaint of an electricity generator that a transmission network has refused to interconnect with it. In contrast, rulemaking concerns the development of new policies because of new technological and market developments. In both cases, adjudication and rulemaking, the regulatory agency can have a fair amount of discretion in its decision-making. This property of regulation is meant to allow the regulators to deal with new and/or contentious issues in a nonpolitical, nonpartisan way. This discretion is limited by the regulatory statute that prescribes what the agency can do and what its guiding purpose is and, very importantly, by due process. The latter is guided by the Administrative Procedures Act. It makes sure that the regulators give all stakeholders the public opportunity to participate in the preparations that lead to the regulatory decisions. This also allows interested parties to challenge the regulatory decisions in court. In fact, courts up to the US Supreme Court have had immense influence on US industry-specific regulation.

Regulators exercise control over the regulated firms via behavioral or structural interventions. Behavioral interventions are usually less drastic. They primarily concern the prices and profits the regulated firms can achieve, where the main control variable in the past has been the allowed rate of return on the firm's invested capital. Another important area of regulatory control

concerns entry, exit and investment. Firms often may not simply start a regulated business and those already under regulation may not simply leave it. Very often, regulated firms have to share their investment plans with the regulator and have them approved. A third area of regulatory concern is that of quality of service. This often takes the form of an obligation to serve or of a common carrier obligation. An obligation to serve means that the regulated firm has to hold enough capacity to serve and has to serve everybody. In contrast, a common carrier obligation means that the company has to serve everybody indiscriminately. Thus, a common carrier obligation is compatible with full trains or buses that can only serve a limited number of people, as long as they do not favor some users over others.

Structural regulation concerns, in particular, horizontal and vertical separation of integrated firms. As indicated, these are very drastic interventions, since they split up going concerns. For that reason, instead of true legal vertical separation, accounting separation is often used in order to preserve some of the economies of scope while still achieving resolutions to conflict of interest problems vis-à-vis vertically separated competitors.

While the US regulatory regime evolved in a contentious political and legal process without any advance planning, the EU established planned regulatory approaches in some of the network industries, in telecommunications in particular. These approaches tried to assure a free flow of services across EU member states associated with free entry of firms in the member states. Thus, even without having a central regulatory EU agency, the EU stipulated regulatory rules, within which the national regulatory agencies (NRAs) had to operate. Similar to the US regulatory commissions, the NRAs are now also largely independent of their governments and are guided by industry-specific laws and due process. Industry-specific regulation in Australia and New Zealand has the special feature that the regulators are part of the agency that also guards competition policy. In Australia, it is the Australian Competition and Consumer Commission (ACCC) and in New Zealand the New Zealand Commerce Commission (NZCC).

India and China, as the countries with the World's largest number of wireless connections, have both started from the traditional ministerial-bureaucratic decision-making model but moved from there in different directions (Liu & Jayakar, 2012). Both countries had to deal with interest group problems and international pressures but responded differently. While India embraced private enterprises and followed a traditional regulatory approach, China kept telecommunications carriers under public ownership. In both cases, political pressures remain so that the Indian regulatory set up is not as independent as in the other countries discussed previously. Liu and Jayakar (2012) characterize the Indian approach as incremental, litigious, and influenced by fractious interest groups.

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