3

A THEORY OF ACCESS COMPETITION

ALTHOUGH THIS IS PRIMARILY a historical work, it must begin with a discussion of theory. The argument of the book is that the refusal of competing telephone companies to interconnect gave them a powerful incentive to expand the scope of their networks. That incentive played a crucial role in bringing about universal service as we know it. More generally, the book is about the problem of interconnecting competing networks, and how those relationships of interconnection lead to competitive or monopolistic industry structures. To clarify the historical treatment of those issues, a theoretical framework regarding network competition is outlined.

The chapter begins with a critique of the common assumption that telephone monopoly can be explained by means of supply side efficiencies alone. It shows that from the standpoint of traditional natural monopoly theory, the telephone system has always been an exceptional and seemingly contradictory case. The next four sections sketch out a theoretical alternative to the natural monopoly paradigm that avoids those problems and, it is hoped, sheds new light on the interpretation of the historical events. In essence, it argues that a better understanding of the unique characteristics of telephone competition and monopoly must come from two sources: i) a better definition of the output of networks and ii) a focus on demand side rather than supply side economies. The discussion of theory is intended to be accessible to readers who are not professional economists, while maintaining a level of logical rigor sufficient to satisfy those who are. (It is possible of course that neither audience will be satisfied with the result, but such are the exigencies of interdisciplinary work.)

Natural Monopoly Theory and the Telephone

Economists typically attempt to explain monopoly organization by reference to the theory of natural monopoly. Although that theory is the main conceptual tool available to account for the existence of a monopoly as pervasive and long-lasting as the telephone system, the uneasy fit between

the two has been apparent for more than seventy years. I will begin with an account of the theory and its development over the years, and then cite six reasons why telephone monopoly posed a puzzle within that theoretical framework.

The Development of Natural Monopoly Theory

From the 1870s to the 1930s, business regulation by specialized regulatory commissions gained acceptance by nearly all states. The thinking behind it was the product of a new school of political economy, born in the populist turmoil of the 1880s, which held that in certain industries competition was destructive and inefficient and ought to be superseded by government regulation. In their attempt to come up with a scientific definition of which industries should be regulated, they developed the concept of natural monopoly.

Natural monopoly theory concentrated on supply-side phenomena; that is, it attempted to explain industrial organization by looking at the costs of the firm. The simplest and most straightforward definition of natural monopoly was articulated in 1887 by Henry Carter Adams, an influential professor who was also the recipient of the first doctorate in Economics awarded by Johns Hopkins University. Adams divided industries into three classes: those with constant returns to scale, those with diminishing returns to scale, and those with increasing returns to scale. Businesses in the first two categories, he believed, could be left to the regulatory pressures of the market. In industries characterized by economies of scale, however, competition was disruptive, inefficient, and temporary. A firm became more efficient as it controlled more of the market. "The control of the state over industries should be coextensive with the application of the law of increasing returns in industries," Adams wrote. [16]

Other theorists concluded that there was no single characteristic defining natural monopoly, though scale economy was always an important factor. Thomas Henry Farrer, the Secretary of the British Board of Trade, listed five separate factors defining inherent monopolies, four of them pertaining to the peculiar fixity of utility infrastructures.[17] The 'natural monopoly' label was coined by Richard T. Ely, a contemporary of Adams's. Ely was a professor of political economy at Johns Hopkins University and the founder of the American Economic Association. Like Farrer, he saw monopoly as the product of a conjunction of factors, including scale economies, a high proportion of fixed to variable costs, and physical obstacles to the multiplication of competing facilities.

Since the time of Ely and Adams, natural monopoly theory, like economic theory generally, has become more refined and formalized. Economists no longer equate natural monopoly with economies of scale as such. In the 1960s James Bonbright contended that a single firm could be the most efficient supplier even when the expansion of output results in increases in average cost.[18] A theoretical breakthrough carne with Faulhaber's (1975) work on the sustainability of cross-subsidies in markets

---

[16] Henry Carter Adams, The Relation of the State to Industrial Action, 1 Publications of the American Economics Association, 465-549 (Jan. 1887).

[17] Farrer's criteria of monopoly were: 1) What they supply is a necessity, 2) They occupy peculiarly favored spots or lines of land, 3) The product or service they supply is used at the place where and in connection with the plant or machinery by which it is supplied, 4) The product or service can be increased in supply without a proportionate increase in plant and capital, 5) The business requires a "certain, and a well-defined harmonious arrangement, which can only be attained by unity." Quoted in Edward D. Lowry, Justification for Regulation: The Case for Natural Monopoly, *Public Utilities Fortnightly* 18-19 (Nov. 8, 1973).

[18] James Bonbright, *Principles of Public Utility Regulation* 14-16 (Columbia University Press 1961).

which were naturally monopolistic.[19] The emergence of the 'contestable markets' school of industrial organization theory, developed by Baumol, Panzar, and Willig, verified Bonbright's observation.[20] In the new theory, cost subadditivity replaced scale economies as the recipe for natural monopoly. Cost subadditivity means that the production costs of one supplier serving all of the market are less than those of any combi- nation of multiple suppliers serving a portion of the market. The improved formalization vindicated Bonbright's earlier observation that a monopoly could be the most efficient supplier in the absence of decreasing costs. At a given output, scale economies are sufficient to make cost functions subadditive, but cost functions can still be subadditive when average costs are increasing.

The revamped industrial organization theory was a powerful advance in that it formalized and mathematicized the definition of natural monopoly. Gone are the clumsy, descriptive lists of special features set out in the works of Ely and Farrer and the early utility textbooks. But the refinement in theory did not change its exclusive focus on supply-side efficiencies. The key to industrial organization was still sought in the way the production costs of the firm(s) responded to changes in the quantity of output. Despite the revolution in analytical technique, the basic conception of natural monopoly, as reflected in the verbal definition, did not change. Natural monopoly was said to exist "when one firm can supply the entire market at less cost than two or more firms."[21]

The Telephone as Natural Monopoly: Six Anomalies

The theory of natural monopoly had developed primarily from observations of the railroad and natural gas industries in the 1880s. The telephone was perceived to be like those industries in that monopoly, once controlled, was thought to possess certain benefits. But if one returns to the writings of the earliest observers of the industry, a very different view of the rationale for telephone monopoly can be found. Instead of pointing to increasing returns or other supply-side efficiencies, the utility economists of the 1920s and 30s asserted explicitly and repeatedly that the telephone had become a monopoly in order to "unify the service."

J. Warren Stehman's *Financial History of AT&T* (1925) was the first comprehensive economic history of the American telephone industry.[22] It was written in the years 1920 to 1922, just as the competitive phase of the industry was drawing to a close. Stehman asserted that "complete monopoly" was "the ideal condition for telephone service," and added that the telephone industry, "perhaps to a greater degree than any other public utility, [is] essentially monopolistic in character." According to Stehman, however, "wasteful duplication of facilities" was not the primary reason for its special status:

---

[19] Gerald Faulhaber, Cross-subsidization: Pricing in Public Enterprise, 65 *American Economics Review* 966 (1975).

[20] William Baumol, John Panzar & Robert Willig, CONTESTABLE MARKETS AND THE THEORY OF INDUSTRY STRUCTURE (Harcourt, Brace, Jovanovich 1982); William Sharkey, THE THEORY OF NATURAL MONOPOLY (Cambridge University Press 1982).

[21] Lowry, *supra* note 17, at 22. Compare Lowry's pre-contestable markets definition with Sharkey's: "There is natural monopoly in a particular market if and only if a single firm can produce the desired output at lower cost than any combination of two or more firms." See also Richard Posner, Natural Monopoly and its Regulation, 21 STANFORD L. REV. 548-643 (Feb. 1969).

[22] J. Warren Stehman, THE FINANCIAL HISTORY OF THE AMERICAN TELEPHONE AND TELEGRAPH COMPANY (Houghton Mifflin, 1925).

> [T]here is an additional and more important peculiarity of the telephone industry: that is, that the efficiency and value of the service depend upon the number of persons with whom the subscriber can communicate. Two telephone systems in a community are a source of great inconvenience and usually of expense to the subscribers. An individual who desires to talk to people on each of the two systems is compelled either to install telephones of both companies or to go, from time to time, to some other place than his residence or place of business to use the telephones of the system to which he is not a subscriber.[23]

The need for universal interconnection was thus recognized as a *separate* and even *stronger* reason than supply-side efficiencies for preventing competition in telephony. Thus, anomaly #1 is that *unification of the service,* not increasing returns on the supply side, was cited by the most informed contemporaries as the reason why a telephone monopoly came about.

Anomaly #2 is even more striking: *those familiar with the telephone industry at the time it became a monopoly believed that it did not possess decreasing costs on the supply side.* On the contrary, the average cost of providing local exchange service was thought to *increase* with the number of subscribers. The main source of the diseconomy was switching technology, specifically, the geometric increase in the number of possible connections as the number of subscribers grew.[24] Within a city, growth in the density of stations could result in decreases in per station expenses, as the additional subscribers led to more efficient utilization of outside plant. But growth in the size of an exchange always increased the average costs associated with switching and maintenance.[25] That generally offset the other economies so that utility commissions usually granted rate increases as exchanges grew. During the 1930s, it was normal for textbooks about public utility regulation to contain explicit discussions of that peculiar aspect of the telephone system. Jones and Bigham's *Principles of Public Utilities,* for example, published in 1931, recognized that subscriber growth produced diseconomies rather than economies. The ultimate justification for monopoly, they maintained, was not scale economies but "the necessity of a unified service."[26] Similar arguments were made in other utility manuals published before 1940.[27] Thus, the cost characteristics of the industry not only failed to conform to the expectations of natural monopoly theory, but actively violated them.

The Jones and Bigham text cited above also dwelt at some length upon another anomaly, even more central than the previous one. In telephone service, the authors observed, it is not obvious what is the appropriate *unit* with which to measure increasing scale. In the early discussions of the diseconomies of scale associated with telephony, economists generally treated the number of subscribers as the measure of the scale of output. But, Jones and Bigham argued, a telephone exchange

---

[23] *Ibid* at 234.

[24] In the manual and electromechanical switches of that period, every terminal was hardwired to every other terminal in the switch, so that as subscribers were added to a switch the number of connections multiplied by N(N-1)/2. Larger telephone exchanges thus had higher average costs than smaller ones. For a detailed history of the diseconomies of growth in switching technology, see Milton Mueller, The Switchboard Problem 30 *Technology & Culture* 534-60 (July 1989).

[25] For a quantitative study of those issues see "Cost of Exchange Telephone Service," memo from Joseph P. Davis to Frederick Fish, October 14 1902. AT&T Archives.

[26] Jones & Bigham, *Principles of Public Utilities* (MacMillan 1931).

[27] G. Lloyd Wilson, James M. Herring & Roland B. Eutscher, *Public Utility Industries* (McGraw-Hill 1936); James M. Herring & Gerald C. Gross, *Telecommunications: Economics and Regulation* (McGraw-Hill 1936), 189.

that connected a user to a larger number of other users was offering a distinctly different service, not more of the same service.[28] The volume of traffic was also an important aspect of telephone system output. Perhaps, they speculated, some composite unit such as the "call-mile-minute" could be developed to provide a more scientific measure of the telephone system's output. Although neither the authors nor other utility economists of the period pursued the matter, the question they raised had profound implications. The concept of the scale of output is fundamental to economic analysis. Natural monopoly theory, in both its classical and modern incarnations, hinges on mathematical analysis of the relationship between scalar variables *P (price)* and *Q (quantity).* Yet here was an open confession that economists did not know how to define Q. Thus we are left with anomaly #3: *in telephony, the unit of output is problematical.*

An intuitively plausible definition of the 'scale' of a network is the number of users. That is in fact the definition used most often by classical and contemporary economists. Equating the number of users with the Q scale, however, *has the paradoxical effect of creating an upward-sloping demand curve* (Anomaly #4). In their work on network externalities, for example, Katz and Shapiro (1985) treat the number of users as the output scale of a network, and explicitly state that firms will raise their prices as more subscribers join.[29] While that assumption is an accurate description of how consumers really do value a growing network, it contradicts everything economics tells us about marginal utility and the downward slope of demand curves. That problem was noted by Allen (1988), who went to extraordinary lengths in an attempt to square that anomaly with orthodox economic theory.[30]

By the time of the debate over AT&T divestiture in the late 1970s and early 1980s, the issue of monopoly organization in telephony had been fully absorbed by the supply-side paradigm. The historical basis of telephone monopoly in universal interconnection, and the early doubts about the paradox of diseconomies and the definition of output, had been largely forgotten. Instead, during *United States* v. *AT&T* econometric studies of Bell system cost functions were brandished by both sides in the courtroom. Oddly, (and that is anomaly #5) *empirical studies of the supply side failed to uncover conclusive evidence of scale and scope economies.* It was clear from empirical studies that there were significant economies of density; i.e., that urban areas were cheaper to serve than rural areas. But some of the most comprehensive studies failed to prove the hypothesis that there were economies of scale and scope across all telecommunications services."[31] Other studies, using different statistical techniques and different measures of output, concluded that there were significant economies of scale and scope.[32] Once again, defining output proved to be problematical. In his review of empirical studies of returns to scale in telecommunications, Littlechild (1979) observed that the only obvious scale economies were

[28] "To one who uses electricity, gas, water and street railways it matters not whether he be served by the same company as his friends, but to the user of the telephone it is highly important that he be on the same system with them and with all those with whom he might wish to get in touch... Jones & Bigham (1931) at 89-90.

[29] Michael Katz & Carl Shapiro, Technology Adoption in the Presence of Network Externalities *Journal of Political Economy* 1985.

[30] David Allen, New Telecommunications Services: Network Externalities and Critical Mass 13 *Telecommunications Policy* 257-71 (Sept. 1988).

[31] Melvyn Fuss & Leonard Waverman, The Regulation of Telecommunications in Canada, Technical Report No. 7, Economic Council of Canada, March 1981; David Evans & James Heckman, A Test for Subadditivity of the Cost Function with an Application to the Bell System 74 *American Economic Review* 620 (1984).

[32] Baldev Raj & H.D. Vinod, Bell System scale economies from a randomly varying parameter rrwdel _ J. EcoN. BUSINESS 247-52 (Feb. 1982); J.B. Smith & V. Corbo, Economies of Scale and Economies of Scope in Bell Canada, Working Paper, Department of Economics, Concordia University, Mar. 1979.

in long distance transmission, whereas the least clear pattern of scale economies was in the local exchange.[33] We need not become too deeply embroiled in the complex and highly technical issues raised by those studies to find corroboration for the main point here: the results of studies of supply-side costs have been equivocal, despite the industry's long-term status as a monopoly.

Occasionally, a modern economist resurrected the old puzzles. The most notable example is in Alfred Kahn's classic two volume treatise, *The Economics of Regulation.* In the course of arguing for a definition of natural monopoly as a product of long- run decreasing average costs, Kahn had this to say about the telephone system:

> There are cases of natural monopoly that would seem at first blush not explicable in terms of long-run decreasing costs. [A]s the number of telephone subscribers goes up, the number of possible connections among them grow more rapidly: local exchange service is therefore believed to be subject to increasing, not decreasing unit costs, when the output is the number of subscribers. And yet, it seems clear that this service is a natural monopoly: if there were two telephone systems serving a community, each subscriber would have to have two instruments, two lines into his home, two bills if he wanted to be able to call everyone else. Despite this apparent presence of increasing costs, in short, monopoly is still natural because one company can serve any number of subscribers (for example, all in a community) at lower cost than two.[34]

That passage bears close analysis. Kahn recognized that the requirements of connecting telephone users forces a competitive system to completely duplicate the network of its rival, and that subscribers in such a competitive market would be forced to pay twice for essentially the same service. But for him, the simple observation that one company can interconnect "any number of subscribers ... at lower cost than two" is sufficient for it to qualify as a traditional natural monopoly. The argument appears persuasive and was often cited by others. In reality, it highlights another theoretical anomaly (#6), namely that the efficiencies which are alleged to make telephone service a natural monopoly occur on the *demand* side and not the *supply* side. Contrary to natural monopoly theory, Kahn's rationale for monopoly is entirely independent of the scale of output (if users are taken as the unit of scale); the elimination of the need for duplicate subscriptions occurs whether a telephone system has 100 subscribers or 100 million subscribers. Moreover, the argument proves that a single firm is more efficient not because it makes telephone service cheaper to *produce,* but because it makes telephone service cheaper to *consume* by eliminating the need for duplicate subscriptions.

To recap, the application of industrial organization theory to the telephone system has generated a series of puzzling inconsistencies:

1. Contemporary observers of the monopolization process insisted that its object was to "unify the service" and not to realize supply-side efficiencies;

[33] Stephen C. Littlechild, Elements of Telecommunications Economics (Institute of Electrical Engineers 1979). Ironically, long-distance transmission is precisely where new competition took root, and local exchange service remained largely monopolistic.

[34] Alfred Kahn, *The Economics of Regulation: Principles and Institutions* Vol. 2, 123 (Wiley 1971).

2. The firm's unit costs appeared to increase rather than decrease as the size of the network grew;
3. There was considerable doubt about the proper definition of output;
4. The most common definition of the scale of a network, the number of subscribers, resulted in a paradoxical, upward-sloping demand curve;
5. Empirical studies failed to verify the existence of the supply-side cost characteristics of a monopoly; and
6. The most convincing argument for the efficiency of a single system was based on demand-side rather than supply-side phenomena.

Despite the number and persistence of those issues, few economists have been willing to make an explicit break with the classical natural monopoly paradigm.

The rest of the chapter proposes an alternative conceptual framework for the analysis of network competition, one that resolves these problems. That theoretical framework has two basic elements. One is a redefinition of the output of networks. The other is a focus on demand-side rather than supply-side economies as the critical determinant of market structure. The latter draws on a new branch of economic theory about the *network externality*. Network externality theory developed in the mid-1970s, independently of the natural monopoly tradition. It uses game theory as well as standard economic techniques to model the way one consumer's demand for a product is affected by the behavior of other consumers. Originally applied to understanding telephone demand, it found fruitful application in economic analysis of standardization and new technology adoption as well. The pioneers of that theoretical literature are Rohlfs (1974), David (1985), Arthur (1989), and Farrell and Saloner (1987). Prior to that, however, the theory has not been applied to that period of telephone competition.

Communications Access Networks as Radically Heterogeneous

A key assumption underlying natural monopoly theory, and indeed most economic analysis, is that a firm's output is composed of homogeneous units. Homogeneity means that each unit of q must be the same as any other unit; or, to put it another way, the product remains constant as the amount produced increases or decreases. That assumption seems plausible enough when the product in question is potato chips, electric power, soft drinks, or wheat. It is easy to imagine identical units of such items increasing or decreasing in quantity along a scale *Q*. When the product is communications access, however, the assumption of homogeneity is both false and misleading.

The most important output dimension of an access network is the people and places it connects. From an economic point of view, neither users nor the locations connected are interchangeable; each one is sui generis. Access to New York is not a substitute for access to Chicago. A telephone connection to one's mother is not a substitute for a connection to a phone sex number. Each unit of access represents a separate output. A telephone directory is a gigantic menu, a Sears-Roebuck catalogue listing all the different access services a user can order by punching numbers on the phone. The economic discreteness of those services is demonstrated forcefully whenever a wrong number is dialed. The wrongly dialed party is not a substitute for the desired party; the system has failed to deliver what the customer wants as surely as when a restaurant brings sake and tofu to a table that ordered beer and pizza.

If each unit of access is a different good, the growth of a network involves an enlargement of the product's scope rather than an increase in scale. Economists have made similar arguments before.[35] With one recent exception,[36] however, even economists who explicitly recognize that point tend to ignore or back away from its implications. For the sake of simplicity, they assume that access is homogeneous and get on with the business of normal economic analysis.[37] To do so, however, assumes away the central problem in the economics of network interconnection and competition, as we shall see. Ignoring the heterogeneity of access is understandable (if not entirely justifiable) in an environment of widespread telephone penetration and interconnected competitors. It is particularly troublesome, however, when analyzing early telephone competition, in which differences in the access units supplied by the networks played a crucial role in the contest.

Figure 3.1 is a simple but useful representation of network output. It is a matrix in which each member of the population (A-*n*) is assigned a row and column. Each cell in the matrix represents an access link or connection between a specific pair of users. Each cell is a separate output (Q), and thus has distinct supply characteristics and its own (downward sloping!) demand curve. Any combination of cells represents a distinct output scope. From the supply side, the efficiency of a network depends on how successfully its engineering can realize economies of scope by sharing facilities across cells. Economies of scale are meaningful only *within* one of the cells. From the demand side, the addition of new users to the network creates an economy of scope for existing users. Users obtain additional service capabilities without a proportional increase in their payments for access.

Figure 3.1
MATRIX REPRESENTATION OF NETWORK OUTPUT

	A				
B	Q_{ab}	B			
C	Q_{ac}	Q_{bc}	C		
D	Q_{ad}	Q_{bd}	Q_{cd}	D	
n	Q_{an}	Q_{bn}	Q_{cn}	Q_{dn}	n

[35] Gerald Brock, Telephone Pricing to Promote Universal Service and Economic Freedom, Federal Communications Commission Office of Plans and Policies, Working Paper #18 (1985). A telephone network is described as N(N-1)/2 different products, where N is the number of persons and N(N-1)/2 is the number of potential conversations.

[36] Nicholas Economides & Lawrence J. White, One Way Networks, Two Way Networks, Compatibility and Antitrust, ms., EC-93-14 July 1993. This paper characterizes networks as complementary components. Customers tend to be identified with a particular component (e.g., an access line in telephone service). Service is a composite good. The addition of users to a network creates economies of scope in consumption.

[37] A typical example the testimony of Nina Cornell, former economist for the Federal Communications Commission, in a 1992 court case regarding telephone interconnection in New Zealand. Cornell wrote: "it could be argued that each potential connection from customer A to customer B is in a separate market from a customer's perspective" but later adds that "looking at each potential as a separate market…is commercially unrealistic. Most customers, when offered a choice among several carriers, select a single carrier to supply a group of such potential connections, rather than selecting a separate carrier for each." The testimony goes on to state that "as long as all local exchange providers are interconnected, duplicate access facilities only raise the cost to consumers with no added benefits." Thus, the heterogeneity problem is passed over by assuming that local carriers will be interconnected and hence that competition involves no choice among imperfect substitutes. Brief of Evidence of Nina W. Cornell, p. 9, Clear v. Telecom, Before the High Court of New Zealand, March 1992.

The author is aware of the fact that some deviation from standard usage is involved in that application of the term "economies of scope." Traditionally, economists have considered the joint provision of local and long distance service, and ancillary services such as security alarms or telegraph service, as an example of scope economies in a telephone network.[38] At risk of being repetitive, it is important to stress that I am applying the concept of scope economies to communications networks in a far more thoroughgoing sense than is usual. That framework views *every* pairwise connection between telephone stations as a separate and distinct output. Hence the term *radical heterogeneity*.

Access Competition

I have stressed the heterogeneity of communications access because the concept neatly explains many of the unique features of competition in the supply of communications access. When competing networks are interconnected, it is easy to ignore the heterogeneity of access because the bundle of connections offered by each network appears to be the same. Heterogeneity becomes particularly important and noticeable, however, when competing networks are not interconnected or compatible. That, of course, was the case in the early era of telephone competition.

Access competition occurs when two or more networks supply access services which could be used as substitutes for each other, but do not provide access to each other. In that type of competition, the scope of the networks becomes one of the most important dimensions of rivalry. Each network offers consumers a different bundle of access units. Networks increase their value to consumers by attracting more users or supplying more access than their rival. The competitive process is complex, however, because users face inherently imperfect substitution choices, and the choices one user makes are affected by the choices other users make. That process differs greatly from the type of competition economists normally consider. It is worthwhile to make that distinction in more formal terms.

In the competition models of neoclassical theory, the quantity of a good demanded by society (Q) is divided up among numerous competing firms (q_1, q_2 --- q_J). The output of each firm is assumed to be homogeneous. Once that assumption is made, two corollaries follow: l) each unit produced by the competing firms is a perfect substitute for every other unit; and 2) each supplier's output comprises an additive share of the total output Q that would be produced by a single firm supplying the entire market; thus, $Q = (q_1 + q_2 + ... + q_n)$. An economist interested in industrial organization can then ask whether the amount Q is produced more efficiently under competitive or monopolistic conditions, or whether firm A or firm B has lower costs in producing amount q.

Those assumptions simply do not work when the output represents communications access. Networks are combinations of many different Q's (communications access units). When competition exists, the market is not divided into additive "shares" of a homogeneous quantity Q; instead, different users join different networks. Assuming that the networks are not interconnected, a user who joins one network is not accessible to the users of the other-unless she purchases access from both. A form of rivalry exists, in that users can choose the combination and price they prefer. But the combinations offered are not identical and therefore are imperfect substitutes. Moreover, the "shares" of

[38] More technically, economies of scope tend to mean supply-side cost subadditivity for the special case of orthogonal outputs.

communications access units offered by competing networks do not sum to constant quantities in different conditions.

Figures 3.2 through 3.5 are Venn diagrams illustrating the possible structures of the market for access. (The sets should be interpreted as groups of users, not as representations of geographic territories.) Figure 3.2 represents a monopoly; a single network connects all N users. Figures 3.3, 3.4, and 3.5 represent the three logically conceivable ways in which the market for access could be divided among two unconnected telephone systems, assuming that network 1 attracts some portion *p* of the available users.

In Figure 3.3, networks 1 and 2 attract two separate and mutually exclusive groups of users, representing a scope of s_1 and s_2, respectively. By definition, the two networks offer completely different combinations of communications access units and cannot be used as substitutes. If the users of the two networks are geographically separated and/or have no interest in obtaining access to each other, then figure 3.3 really represents two cases of figure 3.2 above. If not, then the situation in figure 3.3 would rapidly turn into the one represented by figure 3.4, below.

In Figure 3.4, there are D duplicate users who purchase access from both systems, but $D < N$. In that case, the two networks can be used as perfect substitutes only in the supply of access to group D. Overall substitution is still imperfect, as each network has exclusive control of access to a specific group of users. Indeed, the willingness of some users to purchase access from both systems proves that they are not perfect substitutes.

In Figure 3.5, all subscribers purchase access from both systems ($D = N$). That alternative, universal duplication, makes the substitution choice perfect but creates an intriguing paradox. To be perfect substitutes, every user must join both competing networks. Readers will recognize that as the situation described by Alfred Kahn earlier in the chapter. Kahn stressed its inefficiency; I want to emphasize its practical impossibility. If all users joined two or more competing networks, any user would be able to access all other users on any one of the networks and therefore would have no incentive to duplicate.[39] That is a paradoxical feature of access competition: the greater the percentage of duplication, the closer the combinations of access units offered by competing networks come to being perfect substitutes; but the closer the networks' sets of users come to being identical, the less need there is for duplication.

Taken together, the diagrams prove that: 1) separate networks or incompatible standards are never perfect substitutes; and 2) access competition almost always looks like the model in Figure 3.4 – some users are exclusive to one of the competing networks or standards, while others, who desire more extensive access, purchase access from both systems; 3) the combinations of access units offered by competing networks do not sum to a constant quantity when the same number of users is divided among competing networks.

[39] Such a situation could only come about if the capabilities or services of the network were so different technically as to make them non-competitors (e.g. voice vs. data). But that leads us back to a situation in which the networks are not substitutes.

FIGURE 3.2
Monopoly (Single network)

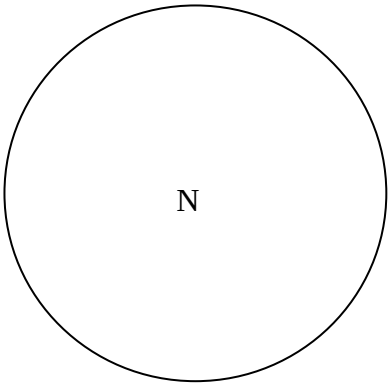


FIGURE 3.3
Dual Networks with no Duplication

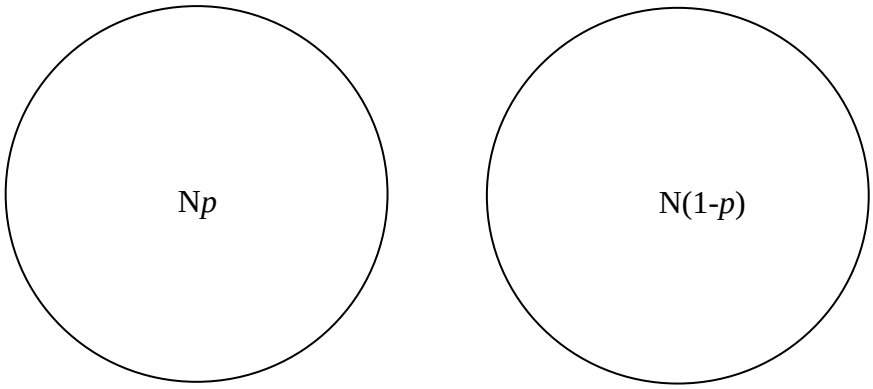


FIGURE 3.4
Competing Networks with Partial duplication

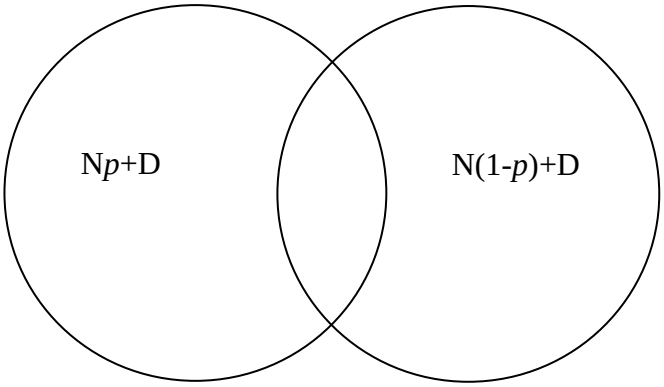
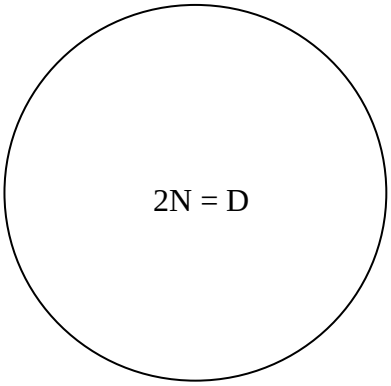


FIGURE 3.5
Dual Networks, Universal Duplication



Imperfect substitution choices give the competitive process a special dynamic. On the demand side, they set in motion a coordination game in which users try to assure themselves of access to all desired parties through joint consumption of the same network. The theoretical literature on network externalities has greatly expanded our understanding of that process, generating a colorful set of terms to describe the unique properties of access competition. Formal models have shown that at any given price for access there can be multiple equilibria. The equilibrium achieved is path dependent; i.e., it can be influenced by the sequence in which users join and other small, random events. There is the problem of achieving the "critical mass" of users required to make joining the network worthwhile. "Bandwagon effects" arise when users who have been "fence sitting" flock to a particular standard or network once critical mass is achieved. There is the danger that users who have committed themselves to a losing standard or network can become technological "orphans." The demand for access and compatibility can also exhibit what Farrell and Saloner call "inertia," or what Arthur and David call "lock-in" effects; users who have converged on a particular network become unwilling to risk sacrificing the benefits of joint consumption by moving to a new network, even when the new alternative is technically more efficient.[40] By making themselves accessible to users of both systems, duplicate subscribers play an important role in stabilizing that process.[41]

On the supply side, access competition puts a premium on universality. Networks with a larger scope are more likely to attract users. More specifically, three incentives to enlarge the scope of a network are created when competing networks are not interconnected:

[40] Theoretical work began with Rohlfs (1974), a game-theoretic model of interdependent demand for communications access. See also Brock (1981); Farrell and Saloner, (1987, 1989); Katz and Shapiro (1985, 1987); David (1985); Arthur (1989); Greenstein (1993). One problem with that literature is its failure to identify the expansion of networks and compatibility as an increase in scope rather than scale. Katz and Shapiro, David, and others erroneously refer to standardization as a product of "demand-side economies of scale." With the exception of Rohlfs, the models tend to treat users as homogeneous and communication patterns as uniform, and thus to overstate the tendencies to converge.

[41] A modified urn model developed by Mueller (1989) showed that convergence on a single network may not happen when there are non-uniform communication probabilities and there is the possibility of duplication by high-volume users.

1) *The incentive to be the first to serve unserved areas or markets.* The inertia associated with joint consumption makes it more difficult to attract existing users away from an established network. New competitors are most likely to gain ground by identifying and attracting new user groups. Thus, access competition is more likely to take place when a market is relatively undeveloped. As a corollary, it is difficult if not impossible to initiate access competition when an incumbent network is near-universal in scope.

2) *The incentive to lower the price of access.* The demand for telecommunications consists of two parts, access and usage. A regime of access competition encourages producers to reduce the cost of, and perhaps even temporarily cross-subsidize, access relative to usage. It also encourages the development of technologies which reduce the cost of access.

3) *The incentive to interconnect users in noncompeting networks.* The quickest way to expand an access universe is to establish connections with an existing network that has already attracted a critical mass of users (assuming, of course, that the existing network is not one's competitor). Competing networks will thus bid for interconnection rights to unaffiliated and noncompeting systems.

All three of those incentives are clearly visible in the historical data developed in subsequent chapters. Together, those three incentives form the basis of my argument that access competition promoted universal service.

Of course, there are corresponding disadvantages to access competition. It is often a transitory process; someone wins the competition and ends up with a monopoly, posing problems of inertia and regulation. Once a certain level of development has been achieved, the existence of separate networks can restrict rather than expand the scope of the system. Duplicate users may be saddled with significant demand-side diseconomies. The fragmentation can be irritating and inconvenient to users. Choosing one network over the other necessarily involves losing access to some potential communication partners. My intention is not to argue that access competition represents the ideal state of affairs. It is, rather, the more limited argument that it played an indispensable part in providing telephone companies the impetus to expand their scope, and that incentive bears the major responsibility for the achievement of universal service.

Access Competition and Appropriability

Economists typically frown upon exploitation of exclusive control of access for competitive advantage. They view the leverage derived from control of access as an exercise of monopoly power.[42] Assuming that there are no insurmountable barriers to the duplication of access facilities, however, it is more accurate to say that access competition represents a qualitatively different *kind* of competition rather than a perversion or suppression of competition. In access competition, rivalry takes place over the *scope* of the product, not just its price. Competition on that dimension is not necessarily socially undesirable because widespread scope is one of the most important determinants of a network's social utility.

[42] See John T. Wenders, *The Economics of Telecommunications* 171-90 (Ballinger 1987), where a telephone company's use of its control of local exchange subscribers to exert leverage over the long distance market is described as an abuse of monopoly power. See also Evans & Heckman, 1983.

In the absence of interconnection or compatibility, a network with a superior scope is able to fully appropriate the economic value of its bundle of access units. Connecting rival networks can eliminate or undermine their ability to appropriate the value of their particular combination of access units. Once again the root of the problem is the network externality, or the interdependence of demand. If the value of the network increases as new users are added, it may be socially efficient to charge some users a price below access costs, and make up the difference by charging higher access rates for users who value the addition of the new users more than the increase in their rates. As Gerald Brock has demonstrated, an access pricing scheme which discriminates among users will be more efficient than one which is uniform, or is based entirely upon cost.[43] A discriminatory pricing scheme which optimizes the scope of a network can only be sustained, however, when free interconnection with a competitor is not required. If interconnection is required, a competitor can undercut the higher access prices and rely on the incumbent to supply access to the users that could only be induced to join the network at a lower price (perhaps even below cost).[44] Thus the incumbent network's ability to appropriate the value of its access bundle deteriorates. The issue of *appropriability* played a major role in the historical drama. Both the Bell and independent telephone interests argued against compulsory interconnection of their networks on those grounds.

Demand-Side Economies of Scope

Understanding the heterogeneity of network output does more than clarify the unique nature of competition among networks; it also improves our understanding of the economic basis for monopoly. The framework established above can be applied to show that imperfect substitution choices can result in user convergence on a single network. The economic gains driving that process come from the demand side rather than the supply side. That framework can also be used to analyze which users have an interest in a monopoly network and who the winners and losers from convergence might be.

As long ago as the 1880s, the promoters of the telephone business remarked that the value of a telephone exchange increased as more people joined it and that the demand for telephone service by one person depended on who else subscribed.[45] That observation, in fact, formed an important part of Theodore Vail's argument for universal service.[46] That insight has been followed up by modern economists, who have given that phenomenon a label ("network externalities") and who have, as noted before, developed formal models of interdependent demand and competition between standards or networks. In what follows, I give that phenomenon a slightly different construction.

The increasing value of networks with a broader scope can be explained as a product of demand-side economies of scope. A user acquires access to a network by buying, building or leasing facilities,

[43] Gerald Brock, *Telecommunications Policy for the Information Age* 72-3 (Harvard University Press, 1994).

[44] *Ibid*. A two-person network connecting A and B charges each $1 for access. Assume that one unit of access costs $1 to supply. A third person, C, is added. Assume that A and B both value access to C at $0.4, and that C values access to A and B at $0.4 each also. C would therefore only be willing to pay $0.8 to join the network. A and B, on the other hand, would be willing to pay up to $1.4. Brock shows that a price vector of $1.3, $1.3, $0.4 will induce all three to subscribe, exactly cover total costs, and make each person better off. If a unit of access costs $1 to supply, however, a competitor could undercut the incumbent's price of $1.3 and offer service to C via interconnection.

[45] George Bartlett Prescott, The Electric Telephone 236 (Appleton 1890).

[46] Theodore N. Vail, AT&T Annual Report 17 (1907); Vail's views are discussed in more detail in chapter 8.

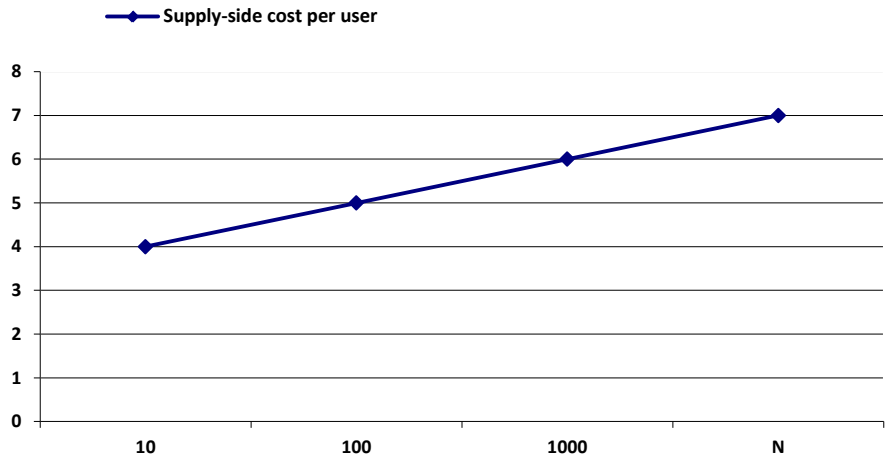
such as a telephone set and a local line. Those investments supply a gateway or entry point into a network, allowing the user to consume a specific set of access services. As additional users join the same network, the number of access services available through those gateway facilities expands. That expansion of service may take place without any increase in the user's investment. Even if the rate paid for access goes up, the increase is likely to be less than what the user would have paid if access to the additional users was purchased separately. Thus, a demand-side economy of scope is realized: additional access units are acquired for a less than proportional increase in user payment.

Conversely, the division of the market into fragmented competing networks can create demand-side diseconomies of scope. Users whose desired calling partners are divided among two or more networks must invest in two or more gateway facilities and subscriptions if they want to maintain access to all of them. Those duplicate investments in access facilities may not be utilized as efficiently as they would be in an integrated system. Returning to the matrix model (Figure 3.1), imagine the costs a user would incur if each pairwise connection, each cell in the matrix, required a separate transaction between the two users involved, a separate pair of instruments, and a separate line. Even with comparatively small networks, the multiplication of access facilities would quickly become monstrously inefficient. Users achieve economies when access units are bundled together.

Integrated networks almost certainly create some supply-side scope economies as well. But demand-side economies of scope can produce efficient user convergence on a single network even when the supply-side costs increase as users are added. That can be illustrated with a simple model (Figure 3.6). In a population of N people, assume the cost per subscriber of supplying telephone service increases as the number of users approaches N. The population is evenly divided among two competing, incompatible networks. Both networks charge $5 per month for telephone service. Under those conditions, a user who wants access to every other user must purchase access to both systems. Thus, universal access costs $10 per month. Now suppose that the city convinces the two systems to consolidate their exchanges into a unified system. The additional costs created by enlarging the integrated system’s scope raise the monthly rate by 20 percent, to $6 per month. Although the rate goes up, the duplicate users have still realized a significant demand- side economy of scope. They now pay $4 less for universal access. Moreover, all users who wanted universal access but were unwilling to pay more than $6 for it have also benefitted from the consolidation.

What was a paradox in natural monopoly theory is now easily explained: one telephone system can be more efficient than two, even when the per-user supply-side cost of one large system exceeds that of two or more smaller, competing systems. The model may make it appear as if a monopoly or a fully interconnected system is *prima facie* more efficient than the alternative. Not so; the realization of demand-side economies of scope in that simple example depended on two assumptions: 1) subscribers had to value access to all other subscribers more than the additional cost created by expanding the scope of the network; and 2) consolidation had to allow duplicate users to reduce the number of access lines they paid for.

FIGURE 3.5
DEMAND-SIDE SCOPE ECONOMIES



Empirically, either one of those assumptions may be untrue. With respect to 1), not everyone wants or needs a system that is universal in scope. Each individual's orders from the 'menu' offered by a universal telecommunications network are different, some being highly extended and others localized and restricted. Under those conditions the elimination of dual service may save money for some groups while raising the costs for many others. The model makes it clear that the distribution of the demand for access among users and the politics of the transition are important empirical is- sues. (Those questions will be explored in chapter 11, when the major urban consolidations of telephone exchanges are examined.) As for assumption 2), large businesses almost always require multiple access lines from the telephone company. Buying access from two competing networks would not necessarily constitute a waste under such circumstances, although it might be an inconvenience due to uncertainty about which one to use to reach specific parties. A company that ordered six access lines under dual service (say, two from one network and four from another) may still need six access lines from a consolidated system. Unless monopoly reduces the number of access lines needed, there is no demand-side economy of scope. (Empirical evidence about subscriber fragmentation and duplication patterns is explored in chapter 7.)

It should also be noted that the existence of a monopoly can restrict the scope of communication as much as, if not more than, the fragmentation caused by competition. The monopoly can charge higher prices for access than it would if faced with competition. It may be unwilling or unable to raise the capital needed to expand as fast as the market demands, or unwilling to risk its money on marginal markets. In general, a system exempt from competitive pressures can be indifferent about increasing the scope of its service.

Interconnection of Competing Networks

Thus far the analysis has assumed that competing networks are not interconnected. To contemporary readers, especially those familiar with current telecommunications policy, that perspective may seem strange, if not downright perverse. Contemporary regulations routinely require open interconnection and equal access. The obvious solution to the problems of access competition, so it would seem, is simply to interconnect the competing networks. That appears to retain the advantages of rivalry while eliminating the problems of imperfect substitution, diseconomies of scope for users, and the danger of eventual convergence on a monopoly. A reader familiar with that modem vantage point will immediately raise two pressing questions about the historical episode: 1) Why didn't public officials mandate interconnection of the competitors rather than permitting access competition to proceed? 2) Why didn't they choose to achieve universal service by interconnecting the independents and Bell, instead of by consolidating the system into a monopoly?

Those empirical questions can only be answered properly in the course of the historical exposition. The issue of how interconnection affects the competitive process is, however, relevant to the theoretical issues raised by that chapter, and are taken up now.

Interconnection homogenizes access. It makes the scope of rival networks appear to users as identical, *even though they are not*. Thus, a firm can offer a substitute for one unit of access without offering a substitute for the entire network. To the customer, the access universe offered is the same, regardless. Users can choose, for example, the local access service of one company and the long-distance service of another. By the same token, a competing network can benefit from the customer access created by a larger network's facilities while invading only those markets that look profitable. Interconnected networks thus have a dual status: they are both complements and competitors. Part of their value is derived from their links to the other network, yet they present themselves to users as substitutes for each other. The long term effects of that process are still unknown, but theory would suggest that it encourages unbundling of the combination of access units making up the network, and discourages rate averaging and cross-subsidization among the units. It also – and that is the critical point – seriously undermines a network's ability to appropriate the value of its scope. A network no longer gains a competitive advantage by maximizing its scope, nor can it maintain that price discrimination that will optimize the scope of the network.

Far from being ignorant of that issue, the telephone companies, users, and municipal and state officials of the early competitive era showed an appreciation of the economic consequences of interconnection that was in many respects more sophisticated than today's reflexive support for it. The main reason access competition persisted was that both competing telephone interests supported it. Their reasoning is described in chapter 5 and chapter 8. Essentially, both wanted to appropriate the value of their networks, and both thought they had a chance to win the competition. Is their attitude any different from the current promoters of incompatible wireless telephone technologies, computer operating systems, or software applications? Clearly, in the developmental stages of a technology, different approaches to compatibility and interconnection seem appropriate. Also, at that period in history, the courts were more willing to accept appropriability-based arguments regarding the property rights of the telephone interests.

Aside from the legal barriers to compulsory interconnection, access competition was often supported or tolerated by the users and public officials because, at that time, access competition was synonymous with competition. Eliminating it via interconnection, they feared, would lead to a state of complementarity between the networks rather than true competition.[47] Access competition was not an accident or a blunder. City councils deliberated for weeks or months before authorizing dual service competition. They were aware of the alternative of a single system. In the later stages of the competitive period, there were also experiments with interconnection of competing exchanges. The experiments tended to confirm the suspicion that competition would cease if rivalry over the scope of the network was eliminated (see chapter 9).

Another important factor was the supply-side cost of interconnection. The network of the early 1900s was not electronic and digital but a mixture of manual and electromechanical analogue. Interconnecting exchanges could not be accomplished automatically, by programming switches, but involved intricate coordination of the procedures of armies of operators. That cooperation would have to take place between business interests with a twenty-year history of hostility and cutthroat competition. Both interests expressed skepticism about the feasibility of such cooperation. Cities balked at its cost in large urban systems. Rather than imposing present-day preconceptions onto the past, that book takes access competition seriously as a historical phenomenon and attempts to explore its characteristics objectively.

To conclude, I have argued that the output of a communication network is radically heterogeneous; i.e., that each connection between users must be considered a separate output, a distinct service. Increases in the number of users attached to a network increase its scope and generally its value to users. Competition over the scope of a network leads to an entirely different kind of business rivalry than competition between firms with outputs that are homogeneous and substitutable. The analysis explored some of the properties of that peculiar form of rivalry and gave it the label access competition. The concept of rivalry over the scope of two non-connected networks provides the theoretical infrastructure for the historical narrative. The Bell-independent rivalry is framed as a history of access competition. Many aspects of the outcome, including the achievement of a ubiquitous telephone infrastructure in the United States, can be attributed to the peculiar incentives generated by competition over the scope of a network. Likewise, the convergence of users on a single network or standard can be seen as a product of demand-side economies of scope.

[47] Stehman, for example, knew that competing companies could be required to interconnect and exchange traffic. But he rejected that as an adequate solution to the problem of service unification. While it eliminated the barriers to communication created by competition, interconnection required the competing companies to make joint financial arrangements and to work so closely together that the result was tantamount to monopoly anyway. Stehman 234 (1925).