Price Theory and Telecommunications Regulation: A Dissenting View

Basil L. Copeland, Jr.†
Alan Severn††

In recent years, analysts have increasingly invoked neoclassical price theory1 to justify radical changes in the regulation of the American telecommunications market, namely the trend toward competition and away from traditional regulatory and rate-setting practices. Many economists assert that competition and marginal-cost pricing2 will eliminate cross-subsidization3 and promote efficient markets for local and long-distance telephone services and telecommunications equipment.4 Price theory, however, does not necessarily support structural reform of the telecommunications industry. Indeed, pre-divestiture pricing policy can be defended on the same grounds its critics employ to advocate structural changes. This article assesses the applicability of static price theory to market conditions currently confronting local operating companies. It challenges the widely held belief that structural reform in telecommunications is necessary to achieve economic efficiency, particularly in intraLATA markets. It recommends that regulators view with skepticism arguments derived from price theory that purport to justify significant departures from historic pricing

† Hess & Lim, Inc.
†† Associate Professor of Finance, School of Business Administration, Temple University

We wish to acknowledge the tireless assistance of the editorial staff of the Yale Journal on Regulation in preparing this article for publication. In addition, we wish to thank Thomas Weiss for his invaluable assistance on matters involving telephone engineering, network design, and regulatory practices in telecommunications.


3. Cross-subsidization often leads to opportunities for inefficient "cream skimming." See infra note 97 and accompanying text.


Copyright © 1985 by the Yale Journal on Regulation.
Part I of this article reviews the historical trend toward competition in telecommunications. Part II demonstrates that price theory can be used to defend pre-divestiture pricing policies. In particular, we argue that flat rates, "cross subsidies," and monopoly suppliers do not necessarily cause inefficiency. In Part III, we maintain that price theory analysis supports the continued regulation of intraLATA telephone service.

I. The Trend Toward Competition in Regulated Telecommunications Markets

To date, the trend in regulatory decision-making toward competition and free entry into once regulated telecommunications markets has largely been confined to interstate markets. For more than a quarter century, technological progress has been the driving force for change in the interstate telecommunications industry, setting into motion a chain of events beginning in 1959 and continuing full-speed today. This technological evolution began with greater opportunities to exploit new communications technologies and innovations in equipment and service offerings. These advances lured regulators into promoting competitive entry in selected segments of interstate telecommunications markets to capture the benefits of innovation. Since competition threatened the financial viability of the regulated network, American Telephone & Telegraph Co. (AT&T) responded by attempting to protect its markets through price changes. Eventually, allegations of predatory pricing and cross-subsidization resulted in the break-up of the Bell System and the policy of deregulation that prevails in the interstate telecommunications industry.

The competitive era in the provision of long-distance telecommunications services began in 1959 with the decision of the Federal Communica-

5. A local access and transport area (LATA) is a geographical unit created after the AT&T divestiture to reflect calling patterns. The local telephone company has a monopoly for both toll and local service within each LATA and is precluded from providing any interLATA telecommunications services. Since most LATAs fall entirely within one state, state utility commissions regulate intraLATA service. Thus, the recent pressure for regulatory reform in this area has been at the state level. Where LATAs cover more than one state, regulation falls within the FCC's jurisdiction. See generally United States v. Western Electric Co., Inc., 569 F. Supp. 990 (D.D.C. 1983).

Recently, state regulatory bodies have come under increasing pressure from Other Common Carriers (OCCs) and resellers to open up intraLATA markets to competition. See, e.g., In Response to Reseller's Petition, Arkansas Agrees to Rehearing on Competition Ban, TELECOMM. REP., Mar. 11, 1985, at 9; MCI, GTE Sprint, and LDTS Appeal W. Virginia Competition Order to State Supreme Court, TELECOMM. REP., Mar. 18, 1985, at 18; Kansas Decides Against IntraLATA Competition; Still Keeps Regulatory Hold on IXCs, TELECOMM. REP., June 17, 1985, at 11. Even the FCC appears to be involved in encouraging intraLATA competition. See Stuart, Ruling on Phone Access Aids Case for Business, N.Y. Times, Nov. 29, 1985, at D1, col. 4.

6. But see supra note 5.
tion Commission (FCC or Commission) in *Above 890'* to allocate a portion of the radio spectrum to high-volume users for private microwave communication systems. Competitive entry in terminal equipment followed in 1968, with the FCC's ruling in *Carter v. AT&T (Carterfone).* In *Carterfone,* Carter Electric Company, a supplier of a device for interconnecting mobile telephone users with other telephone users, challenged tariff restrictions which prohibited the interconnection of telephone devices not owned or supplied by AT&T. The FCC ruled that the tariff restrictions failed to distinguish between devices that would compromise the technical integrity of the system and those that would not. Tariff restrictions against the latter devices were struck down, thereby allowing interconnection of any device that met minimum technical standards.

In the 1970's, the trend toward competition continued with the FCC's approval of the Microwave Communications Inc. (MCI) petition to offer specialized long-distance services. AT&T again attempted to employ its technical integrity argument, as in *Carterfone,* to prevent the entry of MCI into the market for specialized long-distance services. Nevertheless, the FCC allowed entry, believing that specialized carriers would offer innovative and specialized services rather than compete directly with AT&T in the provision of message toll service (MTS) or wide-area toll service (WATS). It was only a matter of time, however, until MCI

---

9. The District Court declined to hear the case pending the conclusion of administrative proceedings, holding that the FCC had primary jurisdiction to resolve all questions relating to the validity of telephone company tariffs. 250 F. Supp. at 192.
11. MCI proposed to supply specialized voice and data services to private-line users that could not be easily accommodated by the existing network. Suppliers of such services were called specialized common carriers (SCCs). *In re Establishment of Policies and Procedures for Consideration of Application to Provide Specialized Common Carrier Services in the Domestic Public Point-to-Point Microwave Radio Service and Proposed Amendments to Parts 21, 43, and 61 of the Commission's Rules,* 29 F.C.C.2d 870 (1971) (First Report and Order), *aff'd sub nom.* Wash. Util. & Transp. Comm'n v. FCC, 513 F.2d 1142 (9th Cir.), *cert. denied,* 423 U.S. 836 (1975).
14. *See AT&T Communications, Wide Area Telecommunications Service, F.C.C. Tariff No. 2, October 16, 1985.* WATS service is designed primarily for business customers who make many long distance calls. The calls are transmitted over the same lines as MTS calls, but the prices are lower than for MTS service.
introduced Execunet service, which competed directly with AT&T’s MTS and WATS services. Overturning an FCC decision prohibiting Execunet, the Court of Appeals in 1978 ordered the FCC to allow MCI to offer long-distance service in competition with AT&T.

Subsequent decisions further enlarged the scope of competition in the long-distance market. Shortly after the Execunet decision, the FCC removed tariff restrictions prohibiting the resale and shared use of leased lines. This decision permitted competitors of AT&T to lease channels at discounted rates, such as WATS, and resell them as value-added services, thus competing directly with AT&T’s more expensive service offerings, such as MTS. Furthermore, the Commission freed these competitors (the Other Common Carriers or OCCs) from rigid price, entry, and service regulations, citing their relatively limited market power.

AT&T responded to competitive threats by engaging in allegedly predatory pricing. Soon after the Commission’s decision in Above 890, AT&T filed its TELPAK tariffs which offered discounts for large bundles of private lines. The tariffs were designed to meet the needs of users who might otherwise construct their own microwave transmission systems.

18. This result is what the court intended. Id.
21. This is consistent with the economic theory of limit pricing, which predicts that a monopoly or dominant firm will reduce prices to the extent necessary to deter entry. See F. SCHERER, INDUSTRIAL MARKET STRUCTURE AND ECONOMIC PERFORMANCE 232-52 (2d ed. 1980); Bain, A Note on Pricing in Monopoly and Oligopoly, 39 AM. ECON. REV. 448, 454-63 (1949); and Gaskins, Dynamic Limit Pricing: Optimal Pricing Under Threat of Entry, 3 J. ECON. THEORY 660 (1971).

AT&T apparently perceived the private provision of telecommunications services as a form of competitive entry. It has been noted that the Gaskins limit pricing model is particularly relevant to the telecommunications market because of the high capital investment necessary to provide telecommunications services. Noll, The Future of Telecommunications, in TELECOMMUNICATIONS REGULATION TODAY AND TOMORROW 41, 55 (E. Noam ed. 1983). For an argument that limit pricing should not be construed as predatory pricing, see Note, Telex v. IBM: Monopoly Pricing Under Section 2 of the Sherman Act, 84 YALE L.J. 558 (1975).
AT&T's filings drew harsh attacks from competitors, sparking the long and still unsettled search for a costing methodology which would allow the Commission to evaluate allegations of cross-subsidization.22

 Allegations of cross-subsidization arose again when AT&T filed its Hi-Lo tariffs in response to MCI's initial entry into the market for "specialized" common carrier services.23 Departing from its traditional practice of rate averaging, AT&T proposed to offer a Hi-Lo tariff in which it partially "deaveraged" rates in order to reduce them for private-line services on high-density routes.24 Thus, when threatened by competition, AT&T reacted by petitioning the FCC to block entry; when that strategy failed, the company filed tariffs to match its competitors' offerings.

 These trends toward competition in interstate markets called into question the traditional principles governing the pricing of telecommunication services. One economist after another sought to educate the FCC and the state regulatory commissions on the implications of static price theory25 in testing for the presence of cross-subsidization. Eventually, debate on this topic came to serve as a backdrop for rationalizing a more far-reaching structural change than the mere introduction of selective competition: the breakup of AT&T.26


24. AT&T's pricing policy had traditionally been based on a national average of per-mile costs. Thus, the rate for service between two small rural communities 1000 miles apart was identical to the rate for service between two large urban centers 1000 miles apart, though the cost of service might have been much lower for the high-traffic route. By selectively entering the market only for the lower-cost, high-traffic routes, MCI could undercut the AT&T rates based on systemwide average costs. 2 A. KAHN, THE ECONOMICS OF REGULATION: PRINCIPLES AND INSTITUTIONS 147 (1971). By "deaveraging" its rates, AT&T sought to remain competitive with the OCCs in high-density markets. AT&T Voice Grade/Private Line Serv. (High Density-Low Density), 55 F.C.C.2d 224, 227 (1975).

25. Cost curves and supply-and-demand diagrams, as ordinarily drawn, are dimensionless with respect to time. In economics, this type of analysis is referred to as "comparative statics" or, as here, "static price theory." See R. HEILBRONER & L. THUROW, THE ECONOMIC PROBLEM 110 (6th ed. 1981). "Regulators, legislators, and businessmen have been increasingly influenced by economic theory as academics have come to play a larger role in government and business." Garfinkel & Linhart, The Transition to Local Measured Telephone Service, PUB. UTIL. FORT., Aug. 16, 1979, at 17, 19.

II. Cost Structure and Pricing

The divestiture of AT&T has been heralded as an opportunity to eliminate cross-subsidization and to price services at their marginal costs, yet static price theory fails to prove that traditional pricing of telecommunications services seriously distorts the allocation of resources. Moreover, static price theory limits its concern to efficiency in the allocation of resources and ignores the effects of scale economies and technical change on the overall level of costs. This Part examines the cost structure of the telecommunications industry to evaluate the arguments used in justifying structural reforms.

Historically, the pricing of telecommunications services has been based on a “value of service” concept, in which services are priced according to their value to a defined group of customers, rather than on the strict “cost of service” approach typical of other public utility services. Although elements of the value-of-service approach can be defended in retrospect by static efficiency criteria, value-of-service pricing was probably pursued as a pragmatic response to the insurmountable difficulties of pricing a multiproduct enterprise with complex cost characteristics “at cost.”

Many economists contend that historical pricing practices involve pervasive cross-subsidies. Three aspects of telecommunications pricing have been increasingly attacked as economically unsound: (1) flat rate pricing for local service, (2) class-wide rate averaging, and (3) contributions from toll revenues toward the recovery of access costs. The economists’ attacks imply—with little supporting empirical evidence—that traditional pricing practices have distorted resource allocation. Alfred Kahn contends:

The simple economic principle is that you don’t charge people for benefits they enjoy of which imposes no cost on society. The proper price of each good or service is what it costs society to supply it—not how much people benefit from it.

27. See Kahn, supra note 4, at 149.
29. See 1 A. Kahn, supra note 2, at 155; J. Bonbright, supra note 28, at 89.
30. Pricing “at cost” usually refers to pricing a service at its marginal cost of production. Value-of-service pricing is closely related in effect, if not in intent, to so-called Ramsey pricing, discussed infra at note 82 and accompanying text.
31. See, e.g., Kahn, supra note 4, at 143-44.
32. See infra note 62 and accompanying text.
34. Kahn, A Needed Dose of Competition, 27 Challenge 24, 25 (1984). In a parenthetical statement that immediately follows, Kahn adds:
Actually the prescription is more complicated: prices should be equated to marginal costs, which consumers will then equate to marginal benefits; and when prices equated to marginal costs don’t bring enough revenues to cover total costs, it may be necessary to discriminate
Price Theory

Current pricing reforms rest on the proposition that economic efficiency requires that all services be priced at cost, a premise which departs significantly from the value-of-service concept. Static price theory as used by these economists, however, does not prove that value-of-service pricing creates distortion. The theory omits consideration of too many factors of critical importance in the telecommunications market and ignores the efficiency effects of scale economies and technological change. Joint costs, common costs, economies of scale, and complementarity of production are all present in the production of telecommunications services; all reduce the relevance of static price theory.

The cost conditions underlying supply of telecommunications services are too complex to be reduced to any "simple economic principle" as Kahn has suggested. Considerations of utility favor flat-rate pricing for local service. In addition, recognition that demand for access is derived demand implies that much of the telecommunications product cannot be unbundled and thereby priced separately.

A critical evaluation of the telecommunications market leads us to conclude that the local operating companies within each LATA remain natural monopolies, and that these monopolies may not be sustainable without continued regulation of price and entry. New entrants might serve one segment of the local market at a lower cost than the monopolists, but such entry would raise the overall average cost of telecommunications services to all consumers. In other words, allowing "hit-and-run" entry on the theory that contestability will preserve competitive price structures is short-sighted.

A. In Defense of Flat Rates

Economists object to flat rate pricing primarily because of cross-subsidization and inefficient levels of consumption. When evaluating claims of cross-subsidization, it is important to distinguish the traffic-sensitive/non-traffic-sensitive dichotomy from the fixed costs/variable costs convention. The important difference is that traffic sensitive costs can be recovered by flat-rate pricing without necessarily incurring the misallocations to which economists object.

among customers on the basis of their evaluations of benefits . . . .
Kahn fails to note that marginal cost prices failing to cover total costs is, according to some, the pricing problem in telecommunications.
35. See M. FRIEDMAN, PRICE THEORY 133 (1976).
36. Kahn, supra note 34.
37. Infra note 72 and accompanying text.
Costs incurred in establishing a telecommunications network are typically divided into non-traffic-sensitive (NTS) costs and traffic-sensitive (TS) costs. Generally, local switching and transmission are traffic-sensitive; the investment required to satisfy customer demands depends on the overall volume and pattern of calls. Similarly, design costs of a long-distance toll network are sensitive to the number of calls being made. On the other hand, the facilities needed to connect a customer to a local end office represent NTS costs. NTS equipment is dedicated to a specific customer; the only calls it carries are to and from that customer.3

The pre-divestiture Bell System approached this dichotomy with value of service pricing—flat-rate pricing for local service and per-minute charges for toll usage. Economists criticized this pricing scheme because it did not properly allocate costs between TS and NTS plant. Critics claimed that value-of-service pricing caused some TS costs associated with the local network to be recovered through the monthly flat rate, while usage-sensitive toll rates allegedly contributed to recovery of NTS local loop costs through the separations and settlements process.9 This interpretation of cost allocation and recovery is arbitrary; it confuses usage-sensitive and traffic-sensitive with the fixed and variable costs of economics. Since toll calls use facilities which are TS but constitute fixed costs, such calls should contribute to the recovery of fixed costs. Indeed, usage sensitive toll rates recovered (and continue to recover) a portion of the fixed cost of the local network other than the local loop. Economists were not concerned about this use of usage-sensitive toll rates to recover fixed

38. The essential elements of a telecommunications network are:
(1) inside wiring and customer premises equipment;
(2) a line drop and a cable pair connecting the customer's premises with a local end office;
(3) local switching and transport; and
(4) toll switching and transport.
The line drop and cable pair are dedicated facilities, or NTS, and represent most of the NTS costs incurred by a local network. See Kahn, supra note 4, at 141. These NTS facilities constitute the link between the customer and all other phones on the system by connecting the phone to the local end office. For a discussion of this access cost, see infra note 62 and accompanying text.

39. Baldwin, Telco Managers Must Take the Initiative in Fighting the Bypass Threat, TELEPHONY, Mar. 5, 1984, at 50 (arguing that historical rate structures have led to "economic aberrations" in the pattern of cost recovery).

"Separations is the process by which the investment and expenses of telephone companies are allocated to the interstate and intrastate jurisdictions. Such allocations provide the mechanism by which revenue requirements for interstate and intrastate operations are developed." In re Prescription of Procedures for Separating and Allocating Plant Investment, Operating Expenses, Taxes, and Reserves Between the Intrastate and Interstate Operations of Telephone Companies. Petition of the National Association of Regulatory Utility Commissioners to Amend Part 67 of the Code of Federal Regulations, 80 F.C.C.2d 230 (1980) (Memorandum Opinion and Order). The current separations and settlements system employed by the FCC incorporates the so-called Ozark plan. As compared to former procedures, the Ozark plan increased substantially the amount of joint and common costs borne by long distance services. In re Prescription of Procedures for Separating and Allocating Plant Investment, Operating Expenses, Taxes, and Reserves Between the Intrastate and Interstate Operations of Telephone Companies, 26 F.C.C.2d 247, 247-64 (1970).
costs. Why not argue that the monthly flat rate was intended to contribute to that part of the local network, again, fixed in cost, but not a part of the local loop, used to route toll calls?

Economists generally dislike flat-rate pricing because in theory it encourages inefficient levels of consumption. Flat rates are supposed to be inefficient because a reallocation of consumption would raise total consumer utility more than it would raise total costs. Likewise rates based upon class-wide cost averaging rather than individual marginal costs can produce allocative inefficiencies. These arguments against rate averaging and flat rates, an extreme example of rate averaging, can be illustrated using basic static price theory. In Figure 1, let $S$ be the commodity supply curve, and $D_1$ and $D_2$ be the demand curves of two different individuals (or classes of service). Efficient prices are determined by the intersections of $D_1$ and $D_2$, respectively, with $S$. That is, consumer utility is maximized when $Q^*_1$ is consumed at price $P^*_1$, and $Q^*_2$ is consumed at price $P^*_2$. Suppose, however, that a non-usage sensitive flat rate of $P_A$ is charged over an interval that spans at least from $Q_1$ to $Q_2$. Alternatively, $P_A$ could be derived by rate averaging, classifying costs along the interval $Q^*_1$ to $Q^*_2$ into a single average cost for both consumers.

In either case, the result is the same. The consumer whose demand is $D_1$ is led by a price $P_A$ to consume only $Q_1$, rather than $Q^*_1$. At $Q_1$, the utility to be derived from additional consumption exceeds the marginal cost of the additional consumption. The total utility to be derived from increasing consumption from $Q_1$ to $Q^*_1$ is the triangle $acd$. But at a consumption level of $Q^*_1$ the consumer faces the price of $P^*_1$, so the net gain (consumer surplus) is the smaller triangle $abd$. In the case of the second consumer, a price of $P_A$ leads to overconsumption, that is, consumption at $Q_2$ rather than $Q^*_2$. If this consumer's consumption were curtailed from $Q_2$ to $Q^*_2$ by changing to price $P^*_2$, the total increase in welfare would be represented by the triangle $egh$. However, this would involve a loss in consumer's surplus given by the triangle $fgh$. The net gain from reducing the second consumer's consumption to $Q^*_2$ is thus only the triangle $efh$.

The following is a typical critique of flat-rate pricing:

Flat rate pricing provides lower total benefits to consumers and producers than a pricing system responsive to the amount of usage consumed. How many pubs provide beer on a flat rate basis? How many supermarkets offer flat rate groceries? How many electric companies offer electricity for a fixed monthly charge?

Flat rate pricing is extremely inefficient when usage varies greatly among customers, when costs vary with usage, when demand is responsive to price, and particularly when the incremental costs of measurement are low.

Consequently, setting prices at cost would increase total consumer welfare by the sum of the triangles $abd$ and $efh$.

Although Figure 1 illustrates the potential for net efficiency gains, the general public may not appreciate such gains. When an economist says that a reallocation of resources will “improve economic efficiency,” the public most likely interprets this to mean lower overall costs because of improvements in productive efficiency. Public approval is unlikely when a regulator says: “On the whole, structural reform may raise your phone bills, but you will feel better about it.” The public does not understand
that economists measure efficiency by consumer utility, not costs. This misunderstanding is never addressed in popular writing on telecommunications reform.\textsuperscript{41}

Figure 1 demonstrates the theoretical static inefficiency of rate averaging but only because we assume that there is a single cost function for serving both types of customers. Without this assumption, a great deal of additional uncertainty surrounds conclusions about the effect of rate averaging on resource allocation. Where the elements of both aggregate demand functions and cost structures are non-homogeneous and uncertain, it is impossible to know what the effect of rate averaging will be. Figure 2, for instance, depicts a case in which rate averaging leads to efficient consumption levels. Suppose there are two customer classes, urban and rural. Assume that the supply costs for rural customers are greater than supply costs for urban customers, but that rural demands are less than urban demands. There might be some average price $P_A$ such that efficient consumption levels result in both markets. While there is some support for the market characteristics hypothesized in Figure 2, the real value of this analysis lies in showing that rate averaging cannot be rejected as inefficient \textit{per se}.

As important as these insights may be, the economic rhetoric against flat-rate pricing and rate averaging is susceptible to a more fundamental critique: TS costs can be efficiently recovered through a flat-rate tariff. The distinction between TS and NTS costs in telecommunications arises not because some costs are fixed and others are not, but because some facilities are dedicated to use by a specific customer. Traffic sensitivity is not synonymous with usage sensitivity. The telecommunications industry stands in sharp contrast to the pub owner.\textsuperscript{42} While the total cost of beer to a pub owner increases every time a patron orders another, the cost of a local telephone network does not increase each time a customer lifts a telephone off the hook.\textsuperscript{43} Thus, the cost of the telephone network is not usage-sensitive, at least until the peak usage exceeds the capacity of the

\textsuperscript{41} The thrust of the popular literature seems to be that elimination of cross-subsidies may drive up local rates, but that this will somehow be offset by a decline in toll rates. But even if cross-subsidies exist and it is possible to improve allocative efficiency by eliminating them, such an elimination does not guarantee a reduction in aggregate costs. An improvement in economic efficiency only requires that consumer utility increase more than aggregate costs. Economists typically avoid accepting responsibility for the distributional consequences of public policy recommendations. It seems to suffice for them to say that those who gain from reallocation could compensate those who lose for the effect of redistributing consumer surplus (the excess of consumer utility over the actual price paid), and that the reallocation is therefore more efficient having, on balance, made everybody better off. But in the real world of public policy, distributional consequences are difficult to ignore. If economists are going to advocate public policy changes, they should at least alert the public to the distributional consequences of their recommendations.

\textsuperscript{42} Crew & Hammelman, \textit{supra} note 40.

\textsuperscript{43} See \textit{infra} note 77.
current network. All plant costs required to render telecommunications service are, strictly speaking, fixed costs, whether or not they are traffic-sensitive. For example, were the TS and NTS distinctions applied in the electric utility industry, all generating capacity would be labelled traffic-sensitive, although its cost is fixed. Absent congestion or a peak-load problem, no economic principles would be violated if this traffic-sensitive generation cost is recovered through a non-usage-sensitive demand charge in a two-part tariff.

It is incorrect, therefore, to insist that static price theory requires recovery of traffic-sensitive telecommunications costs through usage-sensitive pricing. The level of investment required to serve a given volume of traffic is a function of the volume of traffic, but this does not make the cost of traffic-sensitive plant a variable cost as economists ordinarily use the term. Once a given level of capacity is in place, the marginal cost of using that

44. See infra note 48 and accompanying text.
45. Under a two-part tariff, consumers must pay an entry fee calculated to recover overhead costs incurred to provide the service. Each unit consumed is then priced at marginal cost. E. ZAJAC, FAIRNESS OR EFFICIENCY: AN INTRODUCTION TO PUBLIC UTILITY PRICING 37 (1978); I A. KAHN, supra note 2, at 95; and J. BONBRIGHT, supra note 28, at 310.
Price Theory

plant is zero. No economic rationale exists for attempting to limit network use through usage-sensitive pricing. Society gains nothing by inducing consumers to forgo increased use of service if the marginal cost is zero. In static welfare theory, the value of a good or service with zero marginal cost is maximized when each consumer uses the good or service up to the point where the marginal utility of the last unit of use is zero. With a flat rate for unlimited local calling, the level of use selected by customers will presumably reflect the point at which the marginal utility of additional use becomes negative. Customers will use the network up to the point where they would rather be doing something else than chatting on the phone. Although some customers may use the network more than others, in no sense can it be alleged that high-volume users are being subsidized by low-volume users; no cost is avoided if the high-volume users reduce their use of the network.

At some point, of course, the volume of traffic on any given network begins to strain the plant capacity in place. This fact alone, however, does not justify discriminating among customers on the basis of use. Additional investment made to expand the network is a fixed cost that benefits all network users equally. Hence all users, inframarginal or not, cause the costs incurred at the margin. Traffic-sensitive costs incurred to expand the fixed capacity of a network are caused by all users of the network equally and cannot be attributed to one specific class of users. At the margin, all users are responsible for any capacity expansion costs incurred, regardless of whether or not they are the marginal user, since the cost would be avoided if inframarginal users sufficiently reduced their consumption.

A classic example of proper, but counter-intuitive, allocation of incremental capacity costs arises when a water main initially installed to serve a housing subdivision becomes inadequate as additional houses are built. Assume the existing water main must be replaced entirely by a larger

46. Pricing to recover the cost of a telecommunications network is no different conceptually than pricing to recover any large sunk cost such as a bridge or a hydroelectric facility. In these cases, usage-sensitive pricing merely allocates sunk capacity costs in proportion to willingness-to-pay. The person who values the use of an asset most will, presumably, use it the most, and under a usage-sensitive pricing scheme, will then pay the most. However, no additional capacity costs are imposed by those who use the asset more heavily than others, and thus a usage-sensitive tariff or facilities charge cannot be said to be "cost-based" as economists are applying the term in telecommunications. Cf. MacAvoy & Robinson, Losing by Judicial Policymaking: The First Year of the AT&T Divestiture, 2 Yale J. On Reg., 225, 251-52 (1985); Kahn, supra note 4, at 140-41.

47. An exception exists when a network is designed with a certain capacity to meet the demands placed on it during periods of peak use.


49. Contra Kahn, supra note 4, at 147-48. Kahn argues that telephone access costs are analogous to electric distribution costs. We, however, are concerned here with traffic-sensitive costs, which are more appropriately analogous to electric generation or transmission costs.
main. Who should pay for this work? While one might logically argue that the new homeowners created the need for an extension of the main (just as the original homeowners created the need for the original main), economic theory does not provide a logical basis for assigning the responsibility for the entire cost of upgrading the size of the existing main to the new homeowners. The larger size is required to meet the needs of all users, and costs should be assigned accordingly.

Debate over the proper pricing of telecommunications services evidences similar confusion. Consider, for instance, the following argument:

Americans have incorporated the telephone more and more into their way of life, and as the price of substitutes for telephone service (mail and gasoline, for example) have increased rapidly, local use per telephone account has increased. This, of course, has entailed an increase in the cost of providing local service.

The flat rate structure cannot distinguish whether any particular customer has in fact elected to make more use of local service.

To the extent that this argument is based upon usage-sensitive capacity costs, the fact that a customer elects to make more use of the network is irrelevant; any customer could make less use of the network, and additional capacity necessary to supply the increased usage could be avoided. At any given point in time, the marginal cost of using the network is the same for all consumers whether they use or have used the network a little or a lot. Charging rates based upon usage is in reality a form of price discrimination.

The most plausible rationale for usage-sensitive pricing for the local network is the peak-load problem. The existence of a peak period imposes capacity costs that would not be incurred if the volume of use did not vary by the time of day. It is consistent with conventional economic theory to charge those peak-load capacity costs to those who created them. However, this reasoning does not require usage-sensitive rates around the clock. Tariffs with measured service during peak hours and a basic flat rate with unlimited local service at all other times would meet the

50. While the answer may not be in doubt to those existing homeowners who look with disfavor on having to pay any of the cost of upgrading the main, strict economic theory takes a less definite view of the matter. If enough of the original homeowners were to stop consuming, the original main would be adequate to meet the needs of the new homeowners.


52. It may, however, be a rational form of price discrimination. See infra note 82 and accompanying text for a discussion of Ramsey prices or so-called optimal departures from marginal cost. The point here is that charging on the basis of use is not cost-based pricing.

53. When demand varies over time, a network's capacity must be sufficient to meet the maximum demand. A portion of this capacity is idle during all non-peak hours. Thus, capital resources are invested in capacity that is not fully utilized.
standards for economically efficient pricing. Moreover, to the extent that residential calling tends to peak in the early evening, while traffic on the whole peaks during normal business hours, the historic differential between business and residential flat-rates already embodies an element of peak-load pricing. Finally, capacity costs vary according to certain patterns of usage—time of day, duration of call, and distance over which the call is routed. Tariffs incorporating such cost elements serve merely to allocate capacity costs in the same manner in which they are incurred. If the incremental capacity costs are allocated to peak period users, flat-rate pricing with unlimited off-peak use is perfectly consistent with static price theory.

Flat-rate pricing for local service, therefore, is unlikely to cause serious distortions in consumption. If congestion at the peak is indeed a problem, then the cost of local service should be based on a flat rate that incorporates incremental capacity costs. Alternatively, elements of measured service pricing can be used to allocate capacity costs among peak period users with a low, fixed rate for off-peak usage. In any event, the problem of peak-load pricing is hardly unique to telecommunications and provides no rationale for introducing competition into the markets for local telephone service.

B. Joint Costs, Common Costs, and the Issue of Cross-Subsidization

Many economists disfavor flat-rate pricing because it causes high-volume users to subsidize low-volume users, a system which they claim is bound to unravel. Those who make this argument fail to realize that the demand for access to the telephone network is derived from the demand to make local and long-distance calls. This section dispels the myth that unbundling is a panacea for the problem of cross-subsidization.

54. See 1 A. Kahn, supra note 2, at 92-93.
55. The same situation prevails in the pricing of electricity. See Eckel, Customer-Class Price Discrimination by Electric Utilities, 37 J. Econ. & Bus. (forthcoming) (Virg. Polytech. Inst. Dep't of Econ. working paper no. E84-04-01) (on file with the Yale Journal on Regulation). Peak-load pricing, however, has not historically been a rationale for business/residential flat rate differentials. The historical rationale has been the value of service concept.
56. Usage-sensitive toll rates vary by time of day, as they should to respond to the peak-load problem. But this does not justify per se the use of usage-sensitive toll rates. Measured service during the hours of peak use could be combined with a flat rate off-peak, as discussed in the case of local service. To understand why the industry has historically employed usage-sensitive pricing for toll service, one must look beyond strict cost causation.
57. Peak-load pricing is a problem frequently encountered in public utility pricing. See 1 A. Kahn, supra note 2, at 89-103; J. Bonbright, supra note 28, at 359-66. Peak-load pricing is a rate design issue, and thus it does not have a bearing on industry structure.
58. See, e.g., Kahn, supra note 4, at 143.
Rather than attacking flat-rate pricing *per se*, advocates of deregulation and selective competitive entry begin by attacking subsidies which they allege are inherent in the toll revenue settlements process.\(^{59}\) For instance, Alfred Kahn recently maintained that “the inefficiency of our historic method of subsidizing basic service has . . . become intolerable.”\(^{60}\) Similarly, Mark Fowler, Chairman of the FCC and an architect of many structural reforms in telecommunications, has claimed that:

Heavy toll users pay far more than the cost of their loops, while light toll users escape paying part of the cost of their loops. This results in an unintended, poorly targeted cross-subsidy with benefits going to many people who can easily afford to pay the cost of their loops and a corresponding unfair burden shouldered by many residences and small businesses.\(^{61}\)

The cross-subsidy debate focuses upon the tariffs employed to recover the NTS costs of the local network. These access costs\(^{62}\) must be incurred before an individual subscriber can access the network to place a local or toll call. Jules Joskow explains:

Access costs usually are referred to as the non-traffic sensitive costs of providing telephone service. Although these costs are insensitive to usage they, nevertheless, have been recovered through usage charges, primarily on toll service. This means that users of very little long distance service pay substantially less than the cost they impose and high volume users pay substantially, often many times, more than the cost they impose.\(^{63}\)

Joskow here refers to the effect of the separations and settlement process, which historically allocated to local operating companies a portion of the revenues earned from toll services. Under the rationale for the separations process, local operating companies were entitled to at least a portion of the

\(^{59}\) Kahn, *supra* note 4, at 143-44.

\(^{60}\) Kahn, *supra* note 34, at 26.

\(^{61}\) Fowler, *Access Charges Will Do More Good Than Harm*, TELEPHONY, Nov. 21, 1983, at 92. The reference to the “unfair burden shouldered by many residences and small businesses” may be little more than political posturing by Commissioner Fowler, who is astute enough to recognize the emotional issues in the debate over telecommunications policy. Residences and small businesses are as likely to be the light users who, according to Commissioner Fowler, “escape paying part of the cost of their loops.”

\(^{62}\) To access the telephone network, each user needs a connection from her telephone to the local switch. This access, referred to as the local loop, is capital intensive, comprising the most significant of the phone system’s NTS costs. See *supra* note 38 and accompanying text.

revenues from toll rates as compensation for the use of the local network to begin and complete toll calls. 64

It is not necessary for our purposes to examine in any detail the elaborate cost allocation scheme that gradually evolved to facilitate the separations process. We simply note that toll revenues were allocated to local networks on a usage-sensitive basis. 66 Since the cost of the local loop is fixed, this usage-sensitive allocation recovered more of the loop's cost from high-volume users than from low-volume users. This alleged cross-subsidization of low-volume users by high-volume users remains one of the most controversial issues in telecommunications policy. 66

It was once uncontroversial to insist that local and toll service should each make some contribution to the recovery of common or joint costs; the debate was over the amount each service should contribute. In a novel conceptual twist, some economists now seek to sidestep this problem—which has a long history in economics—by unbundling telecommunications into a set of component services. 67 In particular, proponents of unbundling would segregate the cost of service into three components: (1) access or NTS costs; (2) TS local network costs; and (3) TS toll network costs. 68 They would recover the fixed cost of the local loop through a monthly access charge. 69

Conventional economic wisdom posits that fixed costs should be recovered through flat rates and variable costs should be recovered through usage-sensitive rates. 70 This approach does not, however, support the use of fixed charges for the pricing of access, primarily because consumers do not demand access for its own sake. 71

64. The allocation of toll revenues from interstate service to pay the joint costs of interstate and local service was legally sanctioned by the U.S. Supreme Court in Smith v. Illinois Bell Tel. Co., 282 U.S. 133, 150-51 (1930). Since that time, interstate toll revenues have been allocated to pay the costs of connecting customers to the switched network. H.R. REP. No. 479, 98th Cong., 1st Sess. 19 (1983). See supra note 39.


66. Where costs can be unambiguously allocated, subsidization and price discrimination are easy to spot. In telecommunications, however, significant joint or common costs incurred in the provision of local and toll service make direct allocation more problematic.

67. See, e.g., Kahn, supra note 4, at 141, 156-57.

68. Id.

69. Id.; Fowler, supra note 61.

70. See E. ZAJAC, supra note 45, at 37.

71. Admittedly, there is an option good aspect to being connected to the network, but this utility does not derive solely from the cable pair that connects the user with the local end office. A consumer may derive value, for example, from knowing that any important calls can be forwarded from office to home. This requires using elements of the entire network, however, and not just the local loop. Option value is thus not attributable solely to the access line.
The demand for access is a derived demand; access itself possesses no inherent utility to users.\textsuperscript{72} Hence, there is no intrinsic demand curve or demand price for access. Consumers desire access because it is necessary for local or toll calling. We can derive the demand for access from the demand for the ultimate service, but the demand so derived would not exist absent demand for local or long-distance communications.

Unbundling is based on the assumption that toll revenues subsidize access if they contribute any amount to the recovery of access costs. This assumption would be correct if access possessed intrinsic utility to the consumer and the demand for it were not derived. Since demand for access is derived, however, the maximization of static economic efficiency does not depend upon pricing access at cost as an unbundled commodity. Instead, efficient prices for derived goods can be determined by examining the conditions for efficiency in the market for the primary product.\textsuperscript{75}

The classic means of finding the efficient price for a derived good is to subtract from the demand curve of the primary product the supply prices for corresponding amounts of the other factor inputs.\textsuperscript{74} This procedure cannot be applied to access because the demand for access is derived from the demand for two products: local service and toll service. Each of these has an independent utility to the consumer. Thus, it is not clear whether the demand price for access should be derived by subtracting the supply price of local service from the demand for local service or by subtracting the supply price of toll service from the demand for toll service. Indeed, both derived demands must be considered in the access price analysis.

Those who argue that the cost of access should be allocated between local and toll service, somewhat along the lines of the historical separations process, frequently refer to the classic partial equilibrium analysis of joint and common costs.\textsuperscript{76} But that analysis is of only limited usefulness when applied to the access cost issue. The classic analysis of joint or common costs assumes that joint products can be individually priced at their own marginal costs, creating a producer's surplus which covers joint or common costs.\textsuperscript{75} Typically, however, a producer's surplus derives from an upward-sloping supply curve. The cost for any given network capacity, local or toll, is fixed and does not vary with usage. There is, thus, no producer's surplus to fund joint or common costs. These costs must be factored explicitly into the prices to be charged for local and toll service.

\textsuperscript{72} See A. Marshall, Principles of Economics 381-93 (8th ed. 1946).
\textsuperscript{73} See M. Friedman, supra note 35, at 153-56.
\textsuperscript{74} See A. Marshall, supra note 72, at 383.
\textsuperscript{75} See the "blades and handles" analogy in M. Friedman, supra note 35, at 154.
\textsuperscript{76} See W. Sharkey, The Theory of Natural Monopoly 38-39 (1982).
In order to arrive at an efficient pricing scheme, we sum vertically the demand curves for toll and local service to create an aggregate demand curve, as illustrated in Figure 3. The demand curve for local service, $D_l$, and the demand curve for toll service, $D_r$, are summed vertically to produce a joint demand curve, $D_{l+r}$. This aggregate demand curve expresses the consumer's total willingness to pay for toll and local service combined. All costs are treated as fixed and do not vary with use; thus, the system resembles a public good. Aggregate local, toll, and access costs are represented by the cost curve $C_{l+t+a}$, which is hyperbolic. The efficient level of usage from the consumer's perspective is $Q^*$, defined by the intersection of $D_{l+t}$ and $C_{l+t+a}$. Efficient prices for local and toll service, $P_l$ and $P_t$, respectively, can then be derived from the individual demand curves, $D_l$ and $D_r$. Local and toll users each contribute to recovering access costs. The contribution of local service is

$$P_l Q^* - C_l,$$

where $C_l$ is the allocated direct cost of local service. The contribution of toll service is

$$P_t Q^* - C_t,$$

where $C_t$ is the allocated direct cost of toll service.

This analysis has several interesting implications. First, it suggests that the contribution of each class of service to the recovery of access costs is independent of the level of costs that can be directly allocated to either class. Second, and perhaps of more interest, an appropriate allocation of access costs will lead to the same intensity of use in each class of service.

---

77. In other words, we ignore variable costs, both because access costs are fixed and do not vary in the short run with usage, and because telecommunications is intensive with few variable costs. For instance, Howe and Rasmussen report that in 1975 communications had the second lowest fixed asset turnover ratio in the economy, exceeded only by electric, gas, and sanitation services. The fixed asset turnover ratio is a measure of the capital intensity of a business, and is calculated by dividing operating revenues by capital investment. In 1975 the fixed asset turnover ratio for communications was only 0.44, compared to 2.14 for all industries. The fixed asset turnover ratio for all manufacturing, where variable costs play a more significant role than in telecommunications, was 2.31. See K. Howe & E. Rasmussen, Public Utility Economics and Finance 28 (1982). See also 1 A. Kann, supra note 2, at 35. Current estimates of the fixed asset turnover ratio (revenues divided by net plant) for the regional Bell holding companies are: Ameritech, 0.62; Bell Atlantic, 0.58; BellSouth, 0.51; NYNEX, 0.66; Pacific Telesis, 0.55; Southwestern Bell, 0.55; and U.S. West, 0.54. Value Line, Oct. 25, 1985 at 752-84 (Part 3). At the local level, telecommunications continues to be dominated by fixed costs.

78. Public goods are characterized by non-rival consumption and high exclusion costs. Goods which can be consumed by one person without diminishing the supply available for others are nonrival in consumption. Exclusion costs are the costs of limiting consumption to certain people. E. Browning & J. Browning, Microeconomic Theory and Applications 536-37 (1983); R. Musgrave & P. Musgrave, Public Finance in Theory and Practice 56-57 (3d ed. 1980). The telecommunications system is like a public good inasmuch as it is characterized by non-rival consumption, but it is not a pure public good because exclusion can be achieved at minimal cost.
FIGURE 3
Ramsey Efficient Prices For Local and Toll Services

regardless of variations in the level of directly allocable costs. This result derives from our assumption that usage has some of the characteristics of a public good. Since total costs do not vary with use, the efficient level of usage of each service is independent of the directly allocable costs of each service. Thus we have, in effect, a Lindahl equilibrium.

80. By assumption, all costs are fixed and marginal cost equals zero. Usage priced at marginal cost will therefore be free.
81. Lindahl, Just Taxation—A Positive Solution, in CLASSICS IN THE THEORY OF PUBLIC FINANCE 168-76 (R. Musgrave & A. Peacock eds. 1958). Another analyst notes: In a Lindahl equilibrium the role of price and quantity for a public good is exactly the reverse of the role for a private good. Instead of a common price among all consumers, there is a common quantity of consumption. Rather than an allocation of goods among consumers, there is an allocation of the total cost of the public good among all buyers such that the price facing
The telephone network, however, is unlike a pure public good because $Q^*$ is not fixed. Consumption varies and reveals consumer preferences with respect to given levels of consumption. Figure 3 implies that to obtain the equivalent of a Lindahl equilibrium, consumers must not use one type of joint service any more than the other. Suppose that we observe that $Q_t$, minutes of toll use, is greater than $Q_l$, minutes of local use. The implication is that toll service is underpriced relative to local service and is not making a sufficient contribution to the recovery of access costs. A separations procedure which allocates costs in proportion to usage would be self-correcting, and would induce changes in consumption consistent with a Lindahl-type equilibrium. While the settlements process may not have functioned perfectly in this regard, at least it appears to have provided proper incentives for cost allocation.

An attempt to set prices at other than $P_t$ and $P_l$ will produce inefficient resource allocation. Suppose that $P_t$ were reduced to some level $P_t'$, and $P_l$ were increased to some level $P_l'$ by allocating all common costs to local service. Under the assumption in Figure 3 that the elasticity of $D_t$ is less than the elasticity of $D_l$, the revenues lost by moving up the demand curve $D_l$ would be more than offset by revenues gained from moving down the demand curve $D_t$. The result is aggregate revenues in excess of actual costs. Exploiting demand elasticity to recover more than actual costs is price discrimination. Conversely, allocating all common costs to toll service would result in a failure to recover total costs.

The resulting solution is no different than the classical one: the appropriate allocation of common costs is determined by demand conditions in the two markets and is independent of the respective costs of production. Moreover, prices determined by an allocation of costs as suggested in Figure 3 may be Ramsey-efficient. While short-run marginal-cost pricing would set prices at zero, such prices would fail to recover the fixed costs of the business. In a manner similar to the inverse elasticity rule of Ramsey, the Lindahl equilibrium in Figure 3 moves in the direction of the attainable welfare maximum by recovering fixed costs in inverse proportion to the elasticity of demand for the services.  

82. Where pricing at marginal cost fails to recover fixed or common costs, Ramsey pricing will maximize consumer welfare. Ramsey pricing will increase prices over marginal cost. The increase will be smaller in markets where demand is elastic than in markets where demand is inelastic. Thus Ramsey pricing is often dubbed the "inverse elasticity rule." Ramsey prices will reduce each firm's output of goods or services by an equal percentage from the quantity demanded under marginal cost.
Economics teaches us that in a production process with joint or common costs, there is no such thing as pricing individual goods or services at cost. Yet all costs must be recovered. The allocation of access costs between toll and local service on the basis of usage is a pragmatic way to set prices that are as Ramsey-efficient as they can be in practice. On the basis of static efficiency criteria, then, historical pricing practices in telecommunications have not been as inefficient as some critics allege.

C. Contestability, Sustainability, and Public Policy Toward Competitive Entry

In the previous two sections, we have examined two supposed evils of a regulated telecommunications market: flat-rate pricing and cross-subsidization. To eliminate these evils, proponents of deregulation advocate competition as a mechanism for setting prices. In theory, competition will drive prices toward marginal cost, thereby maximizing economic efficiency. Suppliers will therefore no longer be able to set flat-rate prices or subsidize the cost of one product with the revenues from another.

In markets that are natural monopolies, however, competition cannot ensure marginal cost pricing. In natural monopoly markets that are served by a single supplier, only the credible threat of potential competition will guarantee competitive prices. Such markets, where the threat of potential entry is sufficient to maintain competitive equilibrium, are said to be contestable. Deregulation therefore will cure the inefficiencies of flat-rate pricing and cross-subsidization only in those monopoly markets that are contestable.

A market is said to be perfectly contestable if: (1) entry and exit entail minimal sunk investment costs; (2) the technology employed by the incumbent is available to all potential entrants; and (3) the incumbent

pricing. Formally,

\[
e_1 \left( \frac{P^R_1 - MC_1}{P_1} \right) = e_2 \left( \frac{P^R_2 - MC_2}{P_2} \right)
\]

where:

- total revenues = total costs,
- \( P^R_1, P^R_2 \) = Ramsey prices in markets 1 and 2,
- \( MC_1, MC_2 \) = marginal cost prices in markets 1 and 2,
- \( e_1, e_2 \) = elasticity of demand in markets 1 and 2.


84. A successful pricing strategy should maximize consumer welfare. Where joint or common costs exist, this requires attention to class demands. In the broadest sense, this was always the objective of value-of-service pricing. In a rate-regulated multi-product industry with pervasive common or joint costs among product lines, value-of-service pricing is thus an approximation of Ramsey-efficient prices. Usage is a practical surrogate for consumer value.

85. See Kahn, * supra* note 4; MacAvoy & Robinson, * supra* note 46, at 259-60.
cannot reduce prices immediately when faced with the threat of competitive entry. Given these assumptions, the threat of hit-and-run competition will force an incumbent to make welfare-maximizing price and output decisions. Therefore, even a natural monopolist will be forced to refrain from monopolistic pricing—pricing at marginal revenue rather than at marginal cost.

The theory of contestable markets as applied to a natural monopoly differs in only one essential respect from the theory of natural monopoly that has provided the historical framework for public utility regulation. While both posit that under certain circumstances it is desirable to have output supplied by a single firm (the natural monopoly firm), the theory of contestable markets claims that production by a single firm can occur without adverse consequences as a result of market forces while the theory of natural monopoly assumes that it cannot. Absent adverse effects, regulation is unnecessary; potential competition or, more correctly, market contestability guards the public interest by ensuring that the monopolist behaves competitively.

While the theory of contestable markets is full of abstract insights, the conclusions that one draws from it are so dependent upon its assumptions that judging its value and relevance to public policy is difficult. At most, the theory simply emphasizes the importance of determining whether a particular market is indeed a natural monopoly. If a natural monopoly is present, restricting competitive entry is the safest regulatory course. If not, there is no rationale for restricting entry. The theory of contestable


87. In an abstract way, the theory of contestable markets formalizes the notion of workable or potential competition. On workable competition, see F. Scherer, supra note 21, at 41-44.

88. See Snowberger, Sustainability Theory: Its Implications for Governmental Preservation of a Regulated Monopoly, Q. REV. ECON. & BUS., Winter 1978, at 81; Shepard, Subsustainability, Deregulation, and Separate Subsidiaries or Natura Non Facit Saltum, in Challenges for Public Utility Regulation in the 1980s, at 295 (H. Trebing ed. 1981); Shepherd, Contestability v. Competition, 74 AM. ECON. REV. 572 (1984). In the early development of the contestable markets literature, the emphasis was upon sustainability rather than contestability. The theory was thought to justify protection of a monopolist such as AT&T from market entry by competitors such as MCI. Once AT&T negotiated a settlement of the Justice Department antitrust suit, emphasis shifted from sustainability to contestability and the theory became a rationale for allowing entry into regulated telecommunications markets. See, e.g., Testimony of R. Willig, on behalf of AT&T Communications of Maryland, before the Maryland Public Service Commission, Case No. 7788 (Mar. 2, 1984) (on file with the Yale Journal on Regulation). The case for selective competitive entry into long-distance markets predates the development of the contestable markets literature and was a policy promoted by others. See Trebing & Melody, Entry Conditions in Telecommunications, in Regulation and Entry 93 (M. Klass & W. Shepherd eds. 1976). The contestable markets literature appears to have been spawned originally by the search for a rationale to protect the dominant supplier from entry. It is a very flexible theory that can be used to justify policies promoting or precluding entry, depending upon the assumptions made.
markets yields potentially valuable insights only when it is debatable whether markets possess the characteristics of a natural monopoly. The contestability theory has obvious implications for antitrust and regulatory policy. Whether those implications are sufficient to merit regulatory change remains less obvious. The assumption of negligible sunk costs is, of course, critical and open to empirical question in public utility markets. These industries are usually characterized by high fixed costs, many of which are sunk and nonfungible and thus pose significant barriers to entry and exit.

Suppose that a market is a natural monopoly, and it is perfectly contestable. It might then be possible to deregulate the industry without fear of economic harm. The threat of hit-and-run entry would dissipate the market power of the monopolist, rendering traditional rate and entry regulation redundant and perhaps harmful. There is, however, a potential complication. To ensure Pareto optimality (or Ramsey efficiency in the case of a natural monopoly), a monopoly market must not only be contestable, it must also be sustainable. That is, if the monopolist can supply the entire market at the lowest total cost to consumers, there must be no opportunity for an alternative supplier to offer the commodity to some segment of customers at a price below that which the monopolist offers to all customers.

Faulhaber has shown that if average production costs first fall and then rise, as shown in Figure 4, a monopoly is not sustainable. Suppose total demand is equal to $Q_3$. A number of industry structures exist that could satisfy this demand. For example, three firms could each produce $Q_1$, at a unit cost of $P_1$. Since average costs are declining, however, this is not a stable configuration. Alternatively, one firm could produce a level of output equal to $Q_2$, capturing two-thirds of the market, while leaving one-third of the market to be supplied by a single firm producing $Q_1$. A single firm, can, nevertheless, still supply the total market at a lower cost, $P_3$, than any other industry configuration:

$$P_3 < (P_1Q_1 + P_2Q_2)/(Q_1 + Q_2).$$

89. Panzar & Willig, Free Entry and the Sustainability of Natural Monopoly, 8 BELL. J. ECON. 1 (1977).
90. Id.
92. Each firm producing $Q_1$ will attempt to lower its costs and expand its share of the market by expanding to $Q_2$.
93. This expression derives from the relationships depicted in Figure 4 and simply states that the price charged by a single monopoly supplier, $P_3$, is less than the weighted average of the prices paid under competitive entry, $P_1$ and $P_2$. 
In this case, however, total production by a single firm is not sustainable; the monopolist is vulnerable to competitive entry.\footnote{A single firm serving the entire market can be underpriced by an entrant that chooses to supply only part of the market, $Q_2$, at a price of $P_2$. Production of $Q_3$ by a single firm is thus not sustainable in the face of competitive entry.}

FIGURE 4
Contestable Natural Monopoly That Is Not Sustainable

While the sustainability literature has drawn much attention, the notion that a natural monopoly may not be immune to the threat of com-
petitive entry is well known. Faulhaber and others provide new insights into the old problem of cream-skimming.

Economists steeped in neoclassical price theory question the use of the cream-skimming argument to justify regulatory barriers to competitive entry. However, this criticism has tended to assume, either implicitly or explicitly, that costs are separable, that is, all costs of service can be ascribed to a particular customer or class of customers. Where costs are not easily separable, the economists' criticism loses much of its force.

Faulhaber's demonstration of non-sustainability begs a crucial question: why would average costs fall and then rise in the first place? A likely reason is the presence of fixed costs. Fixed costs are common costs of a sort: they are required to serve all customers who take service but are not required to serve any given customer at the margin. They are thus not easily separable, just as common costs are not directly allocable.

Nonseparability of fixed costs is the key to understanding Faulhaber's argument. In Figure 4, average costs are presumed to decline and then rise because of the presence of substantial fixed costs. The fixed costs incurred to supply output $Q_3$, however, are not separable; they are not assignable to any specific customers.

Faulhaber has employed his analysis in the context of prices that are said to be subsidy-free. A price is not subsidy-free if, under any other industry configuration, at least some consumers would enjoy lower prices. A price equal to $P_3$, which requires all consumers to share in the recovery of nonseparable costs, is not subsidy-free, since at least some customers would enjoy a lower price if an entrant captured two-thirds of the market by producing at $Q_2$ and offering price $P_2$. Under this alternative configuration, however, lower prices for some consumers result in higher prices for...
the remaining consumers, and the total cost of service to the market increases. If a competitive entrant captures two-thirds of the market, for example, the output of the monopolist incumbent would fall from $Q_3$ to $Q_1$, and the price paid by the incumbent’s remaining customers would rise to $P_1$. While the result is a stable industry configuration, it is not Pareto optimal or Ramsey efficient. By moving back to a single supplier, the reduction in price from $P_1$ to $P_2$ for the incumbent’s customers would more than offset the rise in price from $P_2$ to $P_3$ for the entrant’s customers.

Faulhaber’s argument can also be explained in another manner. Production by an entrant at $Q_2$, with the former monopolist relegated to producing at $Q_1$, can be interpreted as a form of price discrimination. Demand at any point to the left of curve $D_2$ will be less elastic than demand along curve $D_2$. The entrant who chooses to serve demand $D_2$, while leaving demand $D_1$ to be served by the monopolist, is entering the more elastic of the two markets. The monopolist is forced to discriminate against less elastic customers by raising the price from $P_3$ to $P_1$. This is as much cross-subsidization as the situation posited by Faulhaber, since it hinges on price discrimination against a less elastic market segment.

The inefficient competition analyzed by Faulhaber can occur only if the entrant is not obligated to serve the entire market. Typically, a natural monopolist is granted an exclusive franchise, which requires the monopolist to serve all customers, but prohibits competitive entry. Regulators have feared that eliminating protection from competitive entry would produce wasteful duplication of facilities, loss of revenues from cream-skimming, and higher total costs to consumers. Nonsustainability is merely a more elegant explanation of these intuitions.

Incumbents in a contestable market are presumably unable to change prices immediately. It is therefore assumed that incumbents are unable to respond quickly to the threat of competitive entry. Suppose, however, that regulators allow a natural monopolist to set prices flexibly so that entrants are not given the opportunity to enter the market under the protection of a “price umbrella.” A natural monopolist would then react to the threat of competitive entry by segmenting its markets and pricing $Q_2$ at $P_2$ for the elastic segment of the market and $Q_1$ at $P_1$ for the inelastic segment.

101. The increased consumer costs caused by raising some prices from $P_3$ to $P_1$ more than offsets, in the aggregate, the decrease in consumer costs as the price falls from $P_3$ to $P_2$ for the remainder of the market. In the aggregate, then, total cost to consumers exceeds the total cost that would be collected by an efficient supplier.

102. Faulhaber, supra note 91, at 972. See also Brock & Evans, supra note 97.

Thus, when faced with the threat of entry, the natural monopolist’s defense is to engage in the same kind of inefficient pricing which results from actual entry into a nonsustainable market. To the consumers left paying $P_j$, the result is the same.\textsuperscript{104} The preferable solution is the classic one: grant a franchise to the natural monopolist, require it to serve all customers, regulate prices, and protect it from competitive but inefficient entry.\textsuperscript{106}

The preceding analysis focuses only on the sustainability of a single-product monopoly. Admittedly, the case of a multiproduct monopoly is more complex. But the greater complexity only makes less likely the existence of all the requirements necessary to produce a contestable, sustainable monopoly. The “weak invisible hand”\textsuperscript{108} required to ensure a sustainable set of Ramsey-efficient prices for a multiproduct monopolist will likely be weak indeed.

One serious obstacle to Ramsey pricing in a multiproduct monopoly arises because monopolists are likely to allocate common or joint costs differently than competitive entrants who choose to produce only a subset of the products supplied by the multiproduct monopolist. In such a situation, the monopolist would be forced to engage in price discrimination to prevent competitive entry. Since price discrimination would involve a non-welfare-maximizing pattern of cost allocation or cost recovery, a sustainable set of prices would no longer be optimal. Thus, for multiproduct monopolies, sustainability and optimality may be mutually exclusive.

Whether a given service or set of services is a natural monopoly thus constitutes the threshold question. Contestability alone is insufficient to ensure an outcome that is in the public interest. To obtain a welfare-maximizing result without price and entry regulation, a market must not only be contestable, it must also be sustainable. But if the latter condition is met, we should observe the absence of competitive entry. In a natural monopoly, the existence of competitive entry, especially when it represents entry into only one segment of the market, is likely evidence of non-sustainability. The correct policy response is the traditional one: price and entry regulation.

\textsuperscript{104} The price umbrella argument is thus hardly a defense for allowing flexible pricing by natural monopolies, since, to forestall entry, the monopolist must adopt the same inefficient market segmentation that would result from entry.

\textsuperscript{105} See Brock & Scheinkman, \textit{Free Entry and the Sustainability of Natural Monopoly: Bertrand Revisited by Cournot}, in \textit{BREAKING UP BELL} 231 (D. Evans ed. 1983).

\textsuperscript{106} For a description of the “weak invisible hand”, see Baumol, Bailey & Willig, \textit{supra} note 96, at 350-51.
Price Theory

D. *Static and Dynamic Concepts of Market Efficiency*

In the preceding sections, our justifications of flat-rate pricing and the separations process are based on a static analysis of efficiency in consumption, that is, allocative efficiency. Allocative efficiency concerns the distribution of output to its highest-valued use, not the way output is produced. Efficiency in production, on the other hand, requires that any given quantity of output be produced at minimum cost.

The concept of allocative efficiency is a severely limited framework for policy analysis. It defines efficiency in subjective terms by focusing on the satisfaction of consumer demands rather than on the cost of production. The public, however, seems to expect that structural reform of the telecommunications industry will reduce the cost of telecommunications, as evidenced by the common view that any increase in local rates should be more than offset by a reduction in toll rates. Regulatory reforms leading to a more allocatively efficient rate structure may accomplish a non-quantifiable increase in total consumer welfare, but they do not guarantee reduced telephone prices.

While an increase in allocative efficiency, if attained, would be a worthwhile accomplishment, it is an elusive goal. Furthermore, it is difficult to measure the welfare loss attributable to inefficient allocation induced by entry and price regulation. Attempts to measure the welfare loss from monopoly pricing and output restrictions have met with only limited success and have generally produced such low estimates that some economists have been led to question the usefulness of antitrust actions. It has also been argued that the true cost of monopoly is the lack of diligence on the part of the monopolist in the choice of production techniques, rather than the welfare loss due to monopolistic restriction.

Productive efficiency offers much greater opportunity for increasing consumer utility in the telephone market than allocative efficiency because cost reductions resulting from technological advances and scale economics are enormous. Static price theory takes as given the supply functions and thus ignores the effect of regulation on production efficiency. Joseph Schumpeter argues that the theory of perfect competition is useless as an explanation for the awesome efficiency of capitalist production:

107. See Joskow, *supra* note 63 (discussing the public shock that telephone rates are, on average, rising).
[P]erfectly free entry into a new field may make it impossible to enter it at all. The introduction of new methods of production and new commodities is hardly conceivable with perfect—and perfectly prompt—competition from the start. And this means that the bulk of economic progress is incompatible with it. What we have got to accept is that [capitalism] has come to be the most powerful engine of that progress and in particular of the long-run expansion of total output not only in spite of, but to a considerable extent through, this strategy which looks so restrictive when viewed in the individual case and from the individual point of time. In this respect, perfect competition is not only impossible but inferior, and has no title to being set up as a model of efficiency.

Under certain conditions,111 competitive markets may hinder efficiency in production by stifling technological advances. If a competitive firm’s cost-reducing innovations could be easily and promptly duplicated by its competitors, it would not invest in research and development because it could not raise prices to recover that investment. The same stalemate would occur in perfectly contestable markets.

Allocative efficiency promoted by “correct” pricing policies can have only a minimal effect on what Schumpeter called “the rate of increase of total output,”112 while the exploitation of technological change and economies of scale will have a significant impact on real economic growth. For evidence, one need look no further than what has happened in the telecommunications industry over the last decade. From 1973 through 1981, gross product per person employed in communications increased 5.1 percent per year, whereas for private industry as a whole, the annual rate of increase in gross product per person employed was only 0.4 percent.113 This period presumably coincides with the period during which the “inefficiency of our historic method of subsidizing basic service . . . [became] intolerable.”114 In other words, we are supposed to believe that the most efficient sector of our economy was intolerably inefficient.

111. See F. Scherer, supra note 21, at 423-38. In addition to stifling competition, ruthless price-cutting may pose an obstacle to capital investment in new technology by forcing competitors to base their market response on short-run, rather than long-run, marginal costs. A case in point is the recently deregulated airline industry. Faced with aggressive competition and price cutting, airlines defer maintenance and capital investment that would reduce costs in the long run. Thus, the fleet is aging and airlines are deferring investment in more fuel efficient planes. Paper Reports Airlines Buying More Used Jets, Ark. Gazette, Nov. 18, 1985, at 5A.

112. J. Schumpeter, supra note 110, at 63.


114. Kahn, supra note 34, at 25.
III. Policy Implications

For better or worse, competition in interstate telecommunications is largely a fait accompli. State regulatory bodies are encountering considerable pressure to open up intraLATA markets to competition, and to adopt rate structures that depart radically from historic pricing practices. Federal policy toward telecommunications is the model which state regulators find themselves increasingly pressured to adopt. However, structural differences between interstate, intrastate, and especially intraLATA markets suggest that the kind of competition now common in interstate markets may be inappropriate for intrastate and intraLATA markets.

In our view, historic pricing practices have not caused the degree of price distortion between local and long-distance services that some have alleged. Selective competitive pressures in the interstate and long-haul intrastate markets may be appropriate policy if all suppliers make an appropriate\textsuperscript{115} and equal contribution to the recovery of access costs. The opportunities for cream-skimming are limited because new entrants must employ the same technology and make the same investments as the dominant supplier.\textsuperscript{116} The access charges borne by interLATA long-distance carriers should incorporate not only the cost of accessing the public switched network, but also some portion of the cost of the local loop as well, since both costs are necessary for end-point delivery of long-distance services.

Selective competitive entry is less desirable in intraLATA markets. Local and toll service in proximity to the local network is likely to be a natural monopoly because switching is more likely to be a cost-effective substitute for transmission at the intrastate level than at the interstate level. At the local network level, switching replaces cable pairs running between each and every user. The public switched network is an excellent example of efficient capital substitution: substituting switches and trunk lines for direct cable pairs between each and every user greatly reduces the overall level of capital investment required. The same public switched network which makes local calling possible also routes interLATA calls and intraLATA toll calls to end users. Eventually the public switched network may become obsolete—if the cost of point-to-point communications via satellite, for example, should fall below the cost of a local loop.

\textsuperscript{115} An appropriate contribution to the recovery of access costs is one that requires toll users to make a contribution to the recovery of the cost of the local loop in proportion to usage, as suggested by our analysis of the network as a public good. See supra note 78 and accompanying text. It is the quantity used, rather than the price paid, that characterizes equilibrium between supply and demand in a public goods context, implying that toll users should be allocated local loop costs in proportion to usage.

\textsuperscript{116} Cream-skimming is only possible when a prospective entrant faces lower costs than the incumbent firm. However, the prospective entrant has little room for cost savings when it is required to employ the same technology and capital investment as the incumbent.
But until that happens, the cost of the public switched network will have to be borne by the local operating companies. The presence of this economy of scope, which arises from being able to use the same switched network for local and toll service, suggests that at the local level telecommunications remains a natural monopoly. Since the local telephone company can provide both services efficiently, the local company’s territory should be safeguarded from competitive entry.

The monopoly markets described above are likely to be contestable but nonsustainable.\footnote{See supra notes 86-94 and accompanying text.} In such a market, competitive entry will drive up overall costs without providing attendant social benefits. New entrants will engage in cream-skimming, taking advantage of the existence of the public switched network without having to price at the full cost of the service being provided. More and more of the costs of the switched network will be diverted to incumbent’s remaining customers. If incumbents are granted pricing flexibility to respond to threats of competitive entry, they will engage in the same kind of price discrimination as their rivals practice. While a limited form of price discrimination is consistent with the argument above, it should not be allowed to go so far that toll service customers can escape entirely a ratable allocation of the cost of the local loop. State regulators should therefore limit competitive entry into local telecommunications markets, at least given the presently available technology.

We view the LATAs arising from divestiture as sufficient approximations—for public policy purposes—of the natural monopoly that remains at the local level. They were created with the express purpose of providing the local operating companies with sufficient revenues to remain financially stable and viable. Since the responsibility for the public switched network rests ultimately with the local operating companies, instability and the loss of revenues from intraLATA competition would suggest that those who are using the local network, and benefiting from its existence, are not shouldering their share of the burden. The existence of opportunities for competitors to offer lower prices to selected segments of intraLATA markets is not itself evidence that entry is good public policy. It may simply reflect the fact that the local market, while a natural monopoly, is nonsustainable. But if it is a natural monopoly, the lowest overall cost of service results from prohibiting entry. Until technology advances to the point where the public switched network becomes technologically obsolete, public policy toward telecommunications requires the preservation of financially viable local operating companies, and this in
turn requires a monopoly over intraLATA toll service. It also requires a contribution from interLATA and intraLATA toll users toward recovery of the cost of the local loop. Traditional regulatory objectives, such as universal service; traditional regulatory concerns about cream-skimming and lost revenues if selective entry is permitted; and traditional regulatory mechanisms for achieving these objectives, such as rate averaging and cost allocation, all retain their validity at the local level. Thus, intraLATA service should not be deregulated. The local telephone companies should be allowed to maintain their exclusive franchises to service all calls within each LATA.