Pricing Congestible Network Resources

by

Jeffrey K. MacKie-Mason
Hal R. Varian

University of Michigan

July 1994
Current version: November 17, 1994

Abstract. We describe the basic economic theory of pricing a congestible resource such as an ftp server, a router, a Web site, etc. In particular, we examine the implications of "congestion pricing" as a way to encourage efficient use of network resources. We explore the implications of flat pricing and congestion pricing for capacity expansion in centrally planned, competitive, and monopolistic environments.

Keywords. Networks, congestion, Internet

JEL Classification Numbers. D4, L86, L96

Address. Hal R. Varian, Jeffrey K. MacKie-Mason, Department of Economics, University of Michigan, Ann Arbor, MI 48109-1220. E-mail: `Hal.Varian@umich.edu`, `jmm@umich.edu`

Pricing Congestible Network Resources

Jeffrey K. MacKie-Mason
Hal R. Varian

The Internet is now involved in a major transformation from a government sponsored project to a private enterprise. Privatization and commercialization of the Internet means that providers of network connectivity and services will have to confront issues of pricing and cost recovery. When connectivity was provided to users via government subsidies, little attention was paid to these issues. Suddenly, they have become quite significant. At the same time, new problems in resource allocation are emerging as other telecommunication network technologies begin to converge.

We think that economic modeling can play a significant role in thinking about the consequences of various issues facing decisionmakers. Given the current paucity of economic data about the Internet, economic analysis is unlikely to give precise numerical answers to many questions of interest. Still explicit economic models can serve as a useful guide to "how to think" about some of these issues.

For example, consider the problem of providing bandwidth which will shared by many users. As network technology and availability advances, there will likely be places and periods when bandwidth is scarce and periods when it is abundant. When the supply of bandwidth far exceeds the demand, there is little role for economics. But when the demand for bandwidth exceeds the supply, the fundamental issues of resource allocation become important.

There are many network resources whose performance suffers when there is "overuse": the switching capacity of the routers, the bandwidth of the transport medium, the disk and CPU capacity of popular servers, etc. When users access such resources they presumably take into account their own costs and benefits from usage, but ignore the congestion, delay, or exclusion costs that they impose on other users. Economists refer to this phenomenon as a "congestion externality"; in ecology, it is known as the "problem of the commons" (Hardin (1968)).

There are many ways to deal with congestion externalities. One way is to establish social norms that penalize inappropriate behavior. Such norms can work well in small groups where there is

---

We are grateful to Marvin Sirbu and Scott Shenker their comments. This work was supported by the National Science Foundation grant SES-93-20481.

repeated interaction, but they often do not scale well to a system with millions of users.

Another way to deal with congestion is to establish rationing or quota systems. (Bohn, Braun, Claffy, and Wolff (1993)). One appeal of rationing is that is relatively easy to implement. Indeed, it is common today to see file servers, Web servers, and other network services that reject additional users when the load is too high.

Despite the simplicity of rationing and quotas, economists tend to favor pricing mechanisms as a way of alleviating congestion. One important feature of congestion prices is that they not only discourage usage when congestion is present, but they also generate revenue for capacity expansion. Indeed, it has long been recognized that under certain conditions the optimal congestion prices for a fixed amount of capacity will automatically generate the appropriate amount of revenue to finance capacity expansion.

In previous work we have proposed some simple pricing schemes to deal with congestion (MacKie-Mason and Varian (1993, 1994a)). Here we examine the issue of how the pricing scheme chosen affects industry structure and performance. Our framework is that of "club theory," a term used by Buchanan (1965) to deal with the provision of shared goods. A textbook treatment of club theory can be found in Cornes and Sandler (1986). The papers in the literature that are closest to the treatment here are Scotchmer (1985b, 1985a); we will describe the relationship of our work to this literature in more detail below.

1. Notation

Let x_i denote person i's use of the network resource and $X = \sum_{j=1}^{n} x_j$ the total use of the resource. The user cares about her own use, x_i, and the delay that she encounters. Delay should be interpreted as a general congestion cost: it can include the cost of exclusion, congestion, and so on. Delay depends on the *utilization* of the resource, which we define to be total use divided by capacity: $Y = X/K$. We summarize the preferences of the user by a utility function $u_i(x_i, Y) + m_i$, where m_i is money that the user has to spend on other things. We assume that $u_i(x_i, Y)$ is a differentiable, concave function of x_i and a decreasing concave function of Y.[1]

[1] Later on we consider a special form of this function, $u_i(x_i, Y) = v_i(x_i) - D(Y)$, where $D(Y)$ is interpreted as a delay cost. However, we will not introduce this specification until it is necessary.

The critical feature of this specification is the relationship between usage and capacity: if total usage (X) is doubled and capacity (K) is also doubled, then utilization $Y = X/K$ and hence delay remain constant.[2] We let $c(K)$ measure the cost of providing capacity. For simplicity we take this to be the only cost of providing the service.[3]

This specification is general enough to capture the essence of many network resources. Consider the specific example of an ftp server. In this context x_i could be the number of bytes transferred to user i, K would be the capacity of the server in terms of how many total bytes it can transfer in a given time period, and X would be total bytes transferred to all users. It is natural to suppose that user i cares about the total amount of material she retrieves and the delay involved in retrieving it. A router is another example. In this case x_i would be the bytes sent to (and/or received from) the router by user i, X would be the total use of the router, and K would be the maximum throughput of the router.

2. Efficient use and capacity

We first examine the efficient pattern of usage given some given capacity K. By definition, the efficient pattern maximizes the sum of benefits minus costs. Denoting aggregate net benefits by $W(K)$ we have:[4]

$$W(K) = \max_{x_j} \sum_{j=1}^{n} u_j(x_j, Y) - c(K). \tag{1}$$

The optimal solution must satisfy the first-order condition

$$\frac{\partial u_i(x_i, Y)}{\partial x_i} = -\frac{1}{K} \sum_{j=1}^{n} \frac{\partial u_j(x_j, Y)}{\partial Y} \tag{2}$$

This says that user i should use the system until the marginal benefit from her usage equals the marginal cost that she imposes on the other users.

[2] Delay is fully determined by average utilization only under certain traffic conditions. More generally delay may depend on peak utilization or the variance of utilization. Generalizing the model to account for such effects is clearly of interest, but is beyond the scope of this paper.

[3] In principle, costs could also depend on the amount of usage (X) and on the number of users (n), but we omit these in order to keep the model simple. Capacity costs are normally the dominant costs for most services of interest to us.

[4] We maximize total benefits minus total costs, without making any particular distributional judgments. We could, of course, allow for lump-sum transfer payments to the agents that reflected such concerns. However, such transfer payments would not modify the form of the solution to the benefit-cost problem considered here.

We can decentralize this solution by defining a "shadow price"

$$p_e = -\frac{1}{K} \sum_{j=1}^{n} \frac{\partial u_j(x_j, Y)}{\partial Y}, \tag{3}$$

which measures the total marginal congestion cost that an increase in x_i imposes on the users; note that this is independent of i. Suppose that consumer i is charged a price p_e for usage. Then she would want to solve the following problem

$$\max_{x_i} \; u_i(x_i, Y) - p_e x_i.$$

The solution to this problem is characterized by

$$\frac{\partial u_i(x_i, Y)}{\partial x_i} + \frac{1}{K} \frac{\partial u_i(x_i, Y)}{\partial Y} = p_e. \tag{4}$$

Referring to the definition of p_e in equation (3), we see that for large n the second term on the left-hand side will be negligible relative to the right-hand side of the equation. For large n this expression is essentially the same as the first-order condition for the social optimum given in (2), and thus the decentralized solution corresponds with the social optimum.

To see this more explicitly, consider the special case where $u_i(x_i, Y) = v_i(x_i) - D(Y)$. Then the social optimum in equation (2) is described by

$$u_i'(x_i) = \frac{n}{K} D'(Y),$$

and the individual optimization in equation (4) is

$$u_i'(x_i) = \frac{n+1}{K} D'(Y).$$

For large n these are virtually the same.

Economists say that the price p_e "internalizes" the externality by making the user face the costs that she imposes on the other users. The point of introducing the shadow price is to emphasize the fact that each user should face (essentially) the *same* price for usage—the sum of the marginal congestion costs that each user imposes on the other users.

Capacity expansion

In the maximization problem (1) we used $W(K)$ to denote the maximum welfare given an arbitrary capacity K. What happens to welfare as we expand capacity? Differentiating (1) with respect to K, we have[5]

$$W'(K) = -\sum_{j=1}^{n} \frac{\partial u_j(x_j, Y)}{\partial Y} \frac{X}{K^2} - c'(K).$$

Using the shadow price defined above, we can write this as

$$W'(K) = p_e \frac{X}{K} - c'(K). \tag{5}$$

From this it follows that $W'(K) > 0$ if and only if $p_e X - c'(K)K > 0$. This means that expanding capacity will increase welfare if and only if the revenue from the congestion fees ($p_e X$) exceeds the value of capacity ($c'(K)K$), where capacity is valued using the marginal cost of capacity.

Hence the shadow price p_e plays a dual role: it provides a measure of the social cost of increased usage for an given capacity, but it *also determines the value of a change in capacity*. The fact that congestion fees send the right economic signals to expand capacity under certain conditions was noted by Mohring and Hartwiz (1962) and Strotz (1965); it takes various forms in the literature and is considered a classic principle of congestion pricing.

3. Pricing in a competitive market

The above discussion describes optimal pricing in a utopian world of welfare maximization. In the brave new world of deregulated, privately-provided information network services we would expect to see provision of network resources by profit-seeking firms. What kind of prices would emerge in such a market environment?

The answer depends on the details of market structure: clearly a monopoly or oligopoly structure will result in different (presumably higher) prices than a competitive market. We begin with the admittedly special case of a competitive market with many independent producers; later we examine monopoly provision.

[5] Note that terms involving $\partial x_j / \partial K$ drop out due to the first-order conditions given in (2). This is an instance of what economists call the envelope theorem. (See Varian (1992)).

We suppose that each producer uses a "two-part tariff" for pricing: a "subscription/attachment" fee of q per user per month, say, plus a usage fee of px_i. A representative producer's profits can then be written as

$$\pi = pX + nq - c(K)$$

Here pX is the revenue collected by usage-sensitive fees, nq is the revenue collected from connection fees, and $c(K)$ is the cost of providing capacity K. This appears to be a natural form for pricing network access and usage. Of course, pure connection pricing, in which $p = 0$, and pure usage pricing, in which $q = 0$, are special cases of this pricing form.

Consumer optimization

The utility maximization problem for consumer i is to choose which network resource to use and how much to use it. We suppose that there are (potentially) many suppliers with possibly different utilization levels. Suppliers with lower levels of utilization can charge more due to the better service they provide. We write the price offerings of a representative supplier with utilization Y as $(p(Y), q(Y))$, where $p(Y)$ is the usage fee and $q(Y)$ is the subscription fee.[6]

The utility maximization problem for a representative consumer now becomes

$$\max_{x_i, Y} \; u_i(x_i, Y) - p(Y)x_i - q(Y).$$

That is, the consumer chooses which provider to use (represented by Y) and how much to use (represented by x_i) For convenience, we assume that the menu of offered prices can be treated as a continuous and differentiable function of Y. The consumer's optimization problem has first-order conditions

$$\begin{aligned} \frac{\partial u_i(x_i, Y)}{\partial x_i} - p(Y) &= 0 \\ \frac{\partial u_i(x_i, Y)}{\partial Y} - p'(Y)x_i - q'(Y) &= 0. \end{aligned} \tag{6}$$

The first equation shows that each user will use the resource until the value of additional usage equals its price. The second equation shows that the consumer's choice of delay satisfies the condition that the marginal utility cost of increased delay must be compensated by a reduced

[6] For simplicity, we assume that each firm offers only one class of service; this can easily be generalized.

expenditure, $p'(Y)x_i + q'(Y)$. Adding this last equation up across consumers gives us an expression that we will use below,

$$p'(Y)X + nq'(Y) = \sum_{j=1}^{n} \frac{\partial u_j(x_j, Y)}{\partial Y}. \tag{7}$$

Producer optimization

A representative producer chooses its capacity K and how much bandwidth to supply to users. We assume that there are many competing producers, each of whom takes the price-quality schedules $(p(Y), q(Y))$ as being outside of its control; i.e., determined by the competitive market.

The profit maximization problem facing a representative producer is to choose X and K to maximize profits given the price-quality schedules available in the market

$$\max_{X,K} \; p(Y)X + nq(Y) - c(K),$$

The first-order conditions are

$$p(Y) + p'(Y)\frac{X}{K} + n\frac{q'(Y)}{K} = 0$$

$$-p'(Y)\left(\frac{X}{K}\right)^2 - nq'(Y)\frac{X}{K^2} = c'(K).$$

Collecting terms we can write:

$$p(Y) + \left[p'(Y)X + nq'(Y)\right]\frac{1}{K} = 0 \tag{8}$$

$$-\left[p'(Y)X + nq'(Y)\right]\frac{X}{K^2} = c'(K). \tag{9}$$

Using equations (6) and (7), we can further simplify these equations to

$$p(Y) = \frac{\partial u_i(x_i, Y)}{\partial x_i} = -\frac{1}{K}\sum_{j=1}^{n} \frac{\partial u_j(x_j, Y)}{\partial Y} \tag{10}$$

$$-Y \sum_{j=1}^{n} \frac{\partial u_j(x_j, Y)}{\partial Y} = c'(K)K \tag{11}$$

Comparing (10) to (2) we see that the competitive price will result in the optimal degree of congestion. By combining (10) and (11) we can write

$$p(Y)X = c'(K)K, \tag{12}$$

which leads to the same rule for optimal capacity that we obtained in equation (5).

In this model a competitive supplier is forced to charge the socially optimal price for the quality of service that he offers. Why is the competitive market price equal to the sum of congestion costs? The term $-(1/K)\sum_j \partial u_j/\partial Y$ is how much the other users of the resource would be willing to pay the provider to *refrain* from selling additional usage. If this is less than the price a user is willing to pay for additional usage, the competitive supplier would want to allow more usage. The producer would stop providing additional usage when the price that a user is willing to pay for additional use is balanced with the amount that the other users are willing to pay for a reduction in total usage.

Free entry

If there are no restrictions on entry, firms will enter the industry until profits are driven to zero:

$$p(Y)X + nq(Y) - c(K) = 0.$$

Substituting the expression for $p(Y)$ derived above, we can write the zero-profit condition as

$$nq(Y) = c(K) - c'(K)K.$$

Dividing through by $c(K)$ we have

$$\frac{nq(Y)}{c(K)} = 1 - \frac{c'(K)}{c(K)/K} = 1 - \frac{1}{e},$$

where e is the *elasticity of scale* (average cost over marginal cost). If the marginal cost of capacity is small relative to the average cost, subscription fees will cover most of the cost of providing the service. If the marginal cost of capacity is large, then usage fees will contribute more to recovering total costs.

Scotchmer (1985b) examines a model of two-part pricing of a congestible resource that has some features in common with the one described above. In her model, congestion depends on the number of users, not the total usage, and the capacity of the club is fixed. (This is natural for the kinds of clubs that motivated her study: golf courses, ski clubs, swimming lanes, etc.; it is less natural in our context.) She considers an oligopolistic model with a finite number of firms and examines the limiting behavior as the number of firms increases. She finds that the connection fee goes to zero as the number of firms is increased. Although this result is derived under the assumption that the technology of each firm has fixed capacity, it appears that something similar to this will occur for the more general technology we consider here.

We should emphasize that we have examined the functioning of an idealized competitive industry—a market structure with many firms each having a small market share. There is a very real question of whether the provision of network services will actually have this industrial structure. A full examination of this issue is beyond the scope of this paper, but it is worth observing that a critical consideration is the nature of the cost function for providing network services. If larger firms can produce at lower unit cost, then a large firm will be able to undercut the price of a small firm and there will a tendency to end up with a single firm dominating the market.[7] We should emphasize that this is only a tendency: antitrust policy, regulation, and other tools can be used to influence the ultimate outcome.

We conjecture that a primary factor in determining the industry structure of digital communication networks will be the ease of interconnection. If it is costly or difficult for networks to interconnect, large providers will have an automatic advantage in competing for customers. If this view is correct, there may well be a case for government policy to help coordinate interconnection agreements so as to ensure a healthy competitive environment.

Customer sorting and multiple qualities of service

Nothing in this model implies that there will be a single "optimal" quality of service offered. If all users were identical then the joint solution of equations (6) would yield a single quality Y^*, and associated prices $(p(Y^*), q(Y^*))$. However, user preferences for most services are often heterogeneous: some users may be very intolerant of delay while others may prefer to wait but pay low prices.

When customers have heterogeneous preferences for quality, social welfare is generally *not* maximized by having a single, "high quality" service or product available. Typically, there will be users who would pprefer to accept lower quality in exchange for a price reduction—they value the quality difference less than they value the other goods and services they can buy with the

[7] This is known as the case of a natural monopoly; see Sharkey (1982) for a detailed treatment.

savings. Competition with free entry will then force each quality level to be priced efficiently. Some suppliers will have low prices and high congestion, while others offer high prices and low congestion.

How does a competitive market arrive at the socially optimal variety of price-quality choices? Suppose that there are two types of user: delay-tolerant and delay-intolerant, but only one "average" quality of service is initially offered by all the firms. When would it pay for a firm to offer a different quality of service than its competitors?

By offering a quality of service optimized for one of the groups, a deviating firm could attract all the customers from that group. If the revenue from this deviation exceeds the cost of providing the new quality, this would increase the deviant firm's profits. If there are no fixed costs to creating different qualities we would expect to see as many different qualities as there are types of consumer preferences.

But what if there are large fixed costs to adding new service qualities? In this case it may well not be profitable for a deviant firm to provide a different quality since the entrant may have trouble extracting sufficient profits to cover its costs. Hence the equilibrium number of firms and variety of qualities of service offered will depend on the fixed costs of creating new qualities of service.[8]

When individual users have heterogeneous preferences

Thus far we have considered what happens if different users have different preferences for the resource. What if a single user has different willingness to pay for a resource when using it for different purposes? For example, a user may place a high value on the e-mail access from her network service provider, but a lower value on the ability to engage in real-time video conferencing. If there were small costs of connecting to more than one service provider, then we might see a "restaurant" equilibrium: various providers offering different service qualities at different prices, with a single consumer using more than one provider for different purposes.

However, there may be significant costs of accessing additional providers. For example, it might require having multiple lines running into the home or office, as we now have with telephone, cable

[8] Another factor that influences the number of firms is the presence of "network externalities." These occur when one consumer's utility of connecting to a network depends positively on the number of other users who are connected to the network. See Katz and Shapiro (1985) and Economides (1994) for an analytical treatment of this effect.

and electric lines. If the costs of having multiple providers for multiple services get high enough, then we might expect to see single providers who offer multiple qualities of service. There has been considerable recent interest in the development of integrated services networks; see Braden, Clark, and Shenker (1994) for a proposed multiple quality-of-service architecture for the Internet. Pricing is likely to be an effective mechanism for allocating different service qualities to appropriate uses, although the type of pricing that emerges will depend crucially on the evolution of the technological infrastructure (MacKie-Mason and Varian (1994b)).

Adding capacity

We saw earlier that the efficient congestion prices send the right signals for capacity expansion. Let us see how this works in a competitive market.

Suppose that a competitive firm must decide whether to add additional capacity ΔK. We consider two scenarios. In the first scenario, the firm contemplates keeping X fixed and simply charging more for improved quality of service due to the reduced delay. The extra amount it can charge user j is:

$$\left[q'(Y) + p'(Y)x_j\right] \frac{dY}{dK} \Delta K.$$

Using equation (6) this becomes

$$-\frac{1}{K} \frac{\partial u_j}{\partial Y} \frac{X}{K} \Delta K.$$

Summing this over all consumers and using equation (10) we have

$$p\frac{X}{K}\Delta K.$$

This will increase profits if the increase in revenue is greater than the cost of capacity expansion:

$$p\frac{X}{K}\Delta K - c'(K)\Delta K = \left[p\frac{X}{K} - c'(K)\right] \Delta K > 0.$$

Comparing this to equation (5) we see that profits will increase if and only if net social benefits increase.

In the second scenario, the firm expands its capacity and keeps its price fixed. In a competitive market it will then attract new customers due to the reduction in delay. In equilibrium this firm must have the same delay as other firms charging the same price. Suppose that in the initial equilibrium

$X/K = Y$. Then the additional usage must satisfy $\Delta X = Y\Delta K$. It follows that the increase in in profit for this firm is given by

$$pY\Delta K - c'(K)\Delta K = \left[p\frac{X}{K} - c'(K)\right]\Delta K.$$

Again we see that capacity expansion is optimal if and only if it increases profits.

4. Equilibrium without usage fees

In this model usage fees play two critical roles—they determine both the efficient level of usage and the efficient level of capacity. However, usage-based pricing itself is expensive—it requires an infrastructure to track usage, prepare bills, and collect revenues. These transactions costs may be substantial, and a general examination of usage-based pricing must compare the benefits from improved resource allocation with the costs of accounting and billing. We do not attempt that exercise here. However, it is of considerable interest to examine how a model might function that has no usage fees, but only attachment/subscription fees.

It is convenient to specialize the model described above to a specific form for utility:[9]

$$u_i(x_i, Y) = v_i(x_i) - D(Y).$$

Here $D(Y)$ is directly identified as the "delay costs" from congestion. We assume that $D(Y)$ is an increasing, differentiable, convex function. This says that the delay costs increase with utilization, and that they increase at an increasing rate. Note that this additive form implies that additional delay does not affect the marginal benefits from usage—an admittedly extreme assumption.

For this form of utility, the equilibrium values of (K^e, Y^e) in the world with usage based pricing can be written as

$$\begin{aligned} v_i'(x_i^e) &= \frac{n}{K^e}D'(Y^e) \\ nD'(Y^e)Y^e &= c'(K^e)K^e. \end{aligned} \tag{13}$$

The conditions are found simply by writing the conditions (10–11) for the special form of the utility function that we have adopted.

[9] We make this choice primarily to simplify the exposition; most of the results can be obtained without it, but with somewhat more effort.

Let us now consider what would happen if only attachment pricing were available. Since access is priced, but there is no price for usage we assume that agent i satiates at some point x_i^a. This determines $X^a = \sum_{j=1}^n x_j^a$.

User i's utility maximization problem for Y is

$$\max_Y \; v_i(x_i^a) - D(Y) - q(Y),$$

which leads to the first-order condition

$$-D'(Y) = q'(Y).$$

Adding up across the consumers gives us

$$nq'(Y) = -nD'(Y). \tag{14}$$

The supplier's profit-maximization problem is

$$\max_K \; nq(Y) - c(K),$$

which has first-order conditions

$$-nq'(Y)\frac{X}{K^2} = c'(K).$$

Combining this with equation (14) we see that the equilibrium solution with no attachment pricing only must satisfy the equilibrium condition

$$nY^a D'(Y^a) = c'(K^a)K^a. \tag{15}$$

Comparing this to equation (13) we see that the form of the equation that determines equilibrium capacity is the same with and without usage-based pricing: in either case the amount of capacity will be determined by the willingness to pay for reduction in delay.

However there is one subtlety: even though the *form* of the equation is the same in both cases, it may be that the equilibrium magnitudes of the relevant variables are different. In particular, it can easily happen that the number of users is different with and without usage-based pricing. We must therefore compare the equilibria under the two different scenarios: when the number of users is the same and when the number of users is different.

The number of users is the same

For fixed X, the equilibrium capacity is determined by

$$nD'(X/K)\frac{X}{K} = c'(K)K. \tag{16}$$

The convexity of $D(Y)$ implies that the left-hand side of this equation is a decreasing function of K. The right-hand side will be an increasing function of K, as long as $c''(K)$ is not too negative. Putting these facts together, we have Figure 1A.[10] Certainly equilibrium usage with a zero usage price, X^a, is larger than the equilibrium usage with a positive usage price, X^e. Decreasing X shifts the $nD'(X/K)X/K$ curve down, so equilibrium capacity with usage-based pricing will be less than the equilibrium capacity without usage-based pricing.

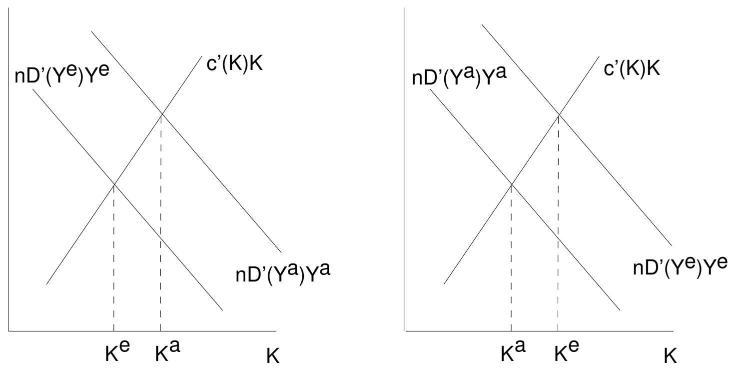

Figure 1. Determination of equilibrium capacity.

Will equilibrium congestion be higher or lower? With zero usage prices each user uses the resource more. But we have just shown that capacity will be higher, too, so it is not obvious what happens to utilization. Consider equation (16) again. Since $D(Y)$ is convex, $nD'(Y)Y$ is increasing in Y. If we write $K = X/Y$, it is easy to see that $c'(X/Y)X/Y$ is decreasing in Y as long as $c(K)$ is convex. Thus we can determine equilibrium congestion as in Figure 2. The increase from X^e to X^a causes $c'(X/Y)X/Y$ to move up, so with no usage pricing there is higher equilibrium congestion.

[10] The curves could be nonlinear; the straight lines are to simplify the presentation.

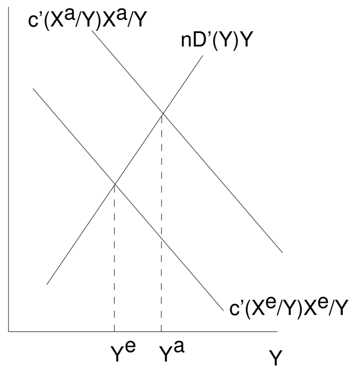

Figure 2. Determination of equilibrium congestion.

The number of users is different

Now we consider the case where the number of users changes. The equilibrium utility of a user without usage-based pricing is

$$v_i(x_i^a) - D(X^a/K^a) - q(Y).$$

This utility could be greater or less than the corresponding utility with usage-based pricing since there is more usage without prices, but there is also more congestion.

Suppose that there is some alternative service that provides the user with utility level u_i^*. Then voluntary participation requires that

$$v_i(x_i^a) - D(X^a/K^a) - q(Y^a) \geq u_i^*,$$

or,

$$v_i(x_i^a) - u_i^* \leq D(X^a/K^a) + q(Y^a).$$

That is, a user will stop using the network under access-only pricing if her net benefit from high usage is less than her congestion cost (including the access fee).[11]

Reducing the number of users will reduce $nD'(Y)Y$. This shifts *down* the corresponding curve in Figure 1B, and could result in an equilibrium amount of capacity that is *less* than one would

[11] We should note that there may also be users who do not consume a usage-priced resource, but *do* consume if there are only access prices. These would be users who want to generate a high volume of low-value traffic.

have under usage-based pricing.[12] One might call this a Yogi Berra equilibrium—after his famous remark that "it's so crowded that no one goes there anymore." In this case, however, the remark is apt: in this equilibrium there are a small number of intensive users with high tolerance for congestion, and therefore low willingness to pay for capacity expansion. The high-value users prefer to exit to alternative services.

5. Market power

What does utilization and capacity look like if there is market power? Suppose, for example, that a resource provider has a monopoly on the resource it provides: e.g., it is the only source for a certain kind of information. In this case it will typically have an incentive to restrict output in order to raise price. How does this affect its choice of optimal capacity?

If the provider prices only on the basis of usage, the answer is pretty straightforward. Generally output will be lower and price higher under a monopoly than under competition. Lower output means that the $nD'(X/K)X/K$ curve will shift down in Figure 1, which implies less capacity.

However, this analysis is based on the assumption of usage pricing only. We have suggested that a combination of attachment and usage pricing would be a fairly common configuration for information and network service providers. The implications of such a two-part tariff are significant.

Identical tastes

For example, suppose that all users have the same tastes. In this case, the maximum connect fee that the monopolist can charge is the fee that makes the user indifferent between using the service and not using it. For simplicity, we normalize the utility of no use to zero, so the participation condition becomes

$$u(x, Y) - px - q = 0.$$

The profit maximization problem of the monopolist is

$$\max_{K,x} \; n[q + p(x, Y)x] - c(K).$$

[12] Reducing n also reduces X^a (which is equal to the sum of satiation usage by all participating consumers), but the convexity of $D(Y)$ ensures that this indirect effect also works to shift $nD'(Y)Y$ downward.

Substituting from the participation condition we have

$$\max_{K,x} \; nu(x, Y) - c(K),$$

which is just the problem of maximizing social welfare. It follows that the optimal policy of the monopolist is to set the use-price equal to the optimal congestion fee, charge the user $q = u(x^e, Y^e) - p^e x^e$ for usage, and make the socially optimal investment in capacity. This observation is the classic two-part tariff result of Oi (1971). See Schmalensee (1981) for a detailed exposition, and Varian (1989) for a survey of this and related results.

Different tastes

However, the assumption that all users—which really means all *potential* users—have identical tastes is rather unrealistic. Let us investigate the more realistic case of heterogeneous users. This case is well-treated in the literature on two-part tariffs cited above, but we need to see how it works for the congestion pricing problem we are examining here.

Let t be a parameter indexing tastes and write the utility function as $u(x, Y, t) = v(x, t) - D(Y)$. Let $f(t)$ be the density of type t and let $F(t)$ be the CDF. Choose the parameterization so that $u(x, Y, t)$ is decreasing in t.

The marginal consumer—the consumer who is just indifferent between using the service or not, denoted by T—is characterized by the condition

$$v(x, T) - D(Y) - q - px(p, T) = 0. \tag{17}$$

For any given p, the monopolist's choice of q is, effectively, a choice of the marginal consumer. Let $X(p, T)$ be the total demand of the consumers who use the service:

$$X(p, T) = \int_0^T x(p, t) f(t)\, dt.$$

The profit maximization problem of the monopolist is

$$\max_{T,p,K} \; qF(T) + pX(p, T) - c(K), \tag{18}$$

where q is defined in (17). Substituting, we have

$$\max_{T,p,K} [v(x, T) - D(Y)]F(T) + p[X(p, T) - x(p, T)F(T)] - c(K).$$

It is worth observing that if the demand of marginal consumer equals the demand of the average consumer, the bracketed term in the middle cancels out and we are back in the previous case.

The first-order conditions for p and K are

$$\left[\frac{\partial v}{\partial x}\frac{\partial x}{\partial p} - D'(Y)\frac{\partial X/\partial p}{K}\right] F(T) + p\left[\frac{\partial X}{\partial p} - \frac{\partial x}{\partial p}F(T)\right] + \\ [X(p,T) - x(p,T)F(T)] = 0 \tag{19}$$

$$D'(Y)YF(T) = c'(K)K.$$

Define the elasticity of demand of the serviced customers as

$$\epsilon = -\frac{p}{X(p,T)}\frac{\partial X(p,T)}{\partial p}.$$

After some manipulations we can write the first-order condition as

$$\frac{p - c'(K)/Y}{p}\epsilon = 1 - \frac{x(p,T)}{X(p,T)/F(T)}.$$

The last term on the right-hand side is the ratio of the demand of the marginal consumer to the demand of the average consumer. If all consumers have the same tastes, then this fraction is 1, and we find that pricing at marginal congestion cost is optimal, as we have already observed. The interesting cases are when the marginal and the average consumer have different tastes.

Recall that by construction the marginal consumer has a lower total value for a given level of usage than the average consumer. Normally, one would think that a consumer with lower total value would want to consume less than a consumer with higher total value. In this case, the monopolist who uses a two-part tariff would set price higher than marginal congestion cost. However, if the marginal consumer wants to consume *more* than the average consumer, it is quite possible that the monopolist would want to set the price *lower* than marginal congestion cost. This is the "auto salesman equilibrium"—the monopolist prices the service so low that he loses money on every sale but makes up for it in volume! Unlike an automobile dealer, our monopolist really may be willing to do this since it collects a uniform subscription fee from each user which can make up for the lost usage fees.

To see how this can happen consider Figure 3, which is based on Oi (1971). There are two users. One has a very high value for the service, but only wants to use a little of it. (Think of ASCII

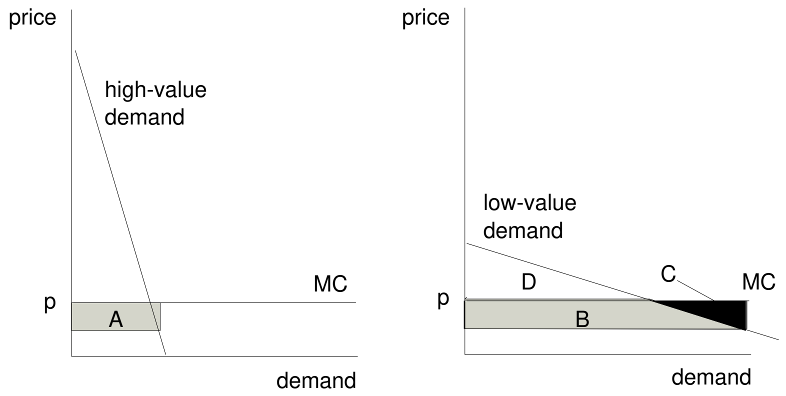

Figure 3. Pricing less than marginal cost may be optimal.

email.) The other user has a low willingness-to-pay for the service but wants to consume a very large amount of it. (Think of a teenager downloading MTV videos.) The teenager is the marginal user, and the connection fee—which is paid by both users—reflects his (relatively low) valuation.

For simplicity, we take the marginal cost of congestion to be constant. Suppose initially that the monopolist prices at the marginal congestion cost; we will show that under some circumstances monopoly profits will increase if the monopolist reduces its price.

If the monopolist sets price equal to marginal cost, the low-value user will achieve net consumer surplus of area D, while the high-value user achieves consumer surplus that is larger than D. The monopolist can therefore charge *each* of them a connection fee of D yielding profits of $2D$.

Now suppose that monopolist reduces its price to some amount *below* marginal cost. The monopolist can now increase the connection fee to $2(D + B)$. However, costs increase as well due to the increased use by both parties. The high-value user imposes additional costs of A and the low-value user imposes additional costs of $(B+C)$. The net increase in profits is $2B-A-(B+C) = B - A - C$. This area may easily be positive, as it is in the case illustrated.

The teenager's utility is larger since he can now download more videos, so he is willing to pay more for the connection; the monopolist extracts this additional surplus through the increased connection charge B. Although the teenager's utility increase (B) is less than the reduction in usage revenues ($B + C$), the email user *also* has to pay the subscription increase (B). In addition the email user will impose costs on the monopolist of an amount A due to his additional usage. Hence if the subscription increase from the high-value user is greater than the usage revenue losses ($B - A - C > 0$), profits will increase when price is set below the marginal congestion cost.

This is the same effect observed by Oi (1971) in his classic article. In the literature it is commonly regarded as a perverse effect that is unlikely to occur in reality. But in our context this effect appears to be quite plausible: it can easily happen that relatively low-valued services can require a huge amounts of bandwidth. In order to capture revenues from such uses, the monopolist may find it profitable to *underprice* the congestion they create, thereby imposing potentially significant congestion costs on high-value, low-bandwidth users.

6. Summary

We have argued that many network resources are congestible: that is, they can be used by more than one person but increasing usage degrades their quality. One person's use creates an externality: it lowers the value of usage for everyone else. Economists long have proposed pricing to internalize this externality: such a price should reflect both the direct and external costs of usage, so that consumers will use the resource efficiently.

In this paper we have developed this theory for a model of the type of congestible resources typically found in an information network. We found that if the resource is provided in a competitive market with connect fees and usage prices, the equilibrium price and capacity will maximize net social benefits. If there is a monopoly provider, however, the profit-maximizing usage price could be either higher or lower than the socially optimal price (with offsetting adjustments in the connection fee), depending on the value that different users put on the resource.

The extent to which the market is competitive ultimately depends on the cost structure of providing the network resource. Whether a given provider will offer a single or multiple qualities of service will depend both on the cost structure and the extent to which an individual user has preferences for multiple qualities of service.

Currently, the most common form of Internet pricing is pricing by access, with no usage-sensitive prices. With a fixed set of users, we expect to see greater capacity when usage is not priced, but also greater congestion. However, with greater congestion, congestion-sensitive users might not use the resource; the resulting "Yogi Berra" equilibrium might actually have lower usage (but higher congestion) than when usage is priced.

References

Bohn, R., Braun, H.-W., Claffy, K., and Wolff, S. (1993). Mitigating the coming Internet crunch: Multiple service levels via precedence. Tech. rep., UCSD, San Diego Supercomputer Center, and NSF.

Braden, R., Clark, D., and Shenker, S. (1994). Integrated services in the Internet architecture: an overview. Tech. rep., IETF. RFC 1633.

Buchanan, J. (1965). An economic theory of clubs. *Economica*, *32*, 1–14.

Cornes, R., and Sandler, T. (1986). *The Theory of Externalities, Public Goods, and Club Goods*. Cambridge University Press, Cambridge.

Economides, N. (1994). Critical mass and network size. Tech. rep., New York University Stern School of Business, New York.

Hardin, G. (1968). The tragedy of the commons. *Science*, *xx*, 1243–47.

Katz, M., and Shapiro, C. (1985). Network externalities, competition and compatibility. *American Economic Review*, *75*, 424–440.

MacKie-Mason, J. K., and Varian, H. (1993). Some economics of the Internet. Tech. rep., University of Michigan. Available from ftp://gopher.econ.lsa.umich.edu/pub/Papers.

MacKie-Mason, J. K., and Varian, H. (1994a). Pricing the Internet. In Kahin, B., and Keller, J. (Eds.), *Public Access to the Internet*. Prentice-Hall, Englewood Cliffs, New Jersey. Available from ftp://gopher.econ.lsa.umich.edu/pub/Papers.

MacKie-Mason, J. K., and Varian, H. R. (1994b). Economic FAQs about the Internet. *Journal of Economic Perspectives*, *8*(3).

Mohring, H., and Hartwiz, M. (1962). *Highway Benefits: An Analytical Approach*. Northwestern University Press, Evanston.

Oi, W. (1971). A Disneyland dilemma: two-part tariffs for a Mickey Mouse monopoly. *Quarterly Journal of Economics*, *85*, 77–96.

Schmalensee, R. (1981). Monopolistic two-part pricing arrangements. *Bell Journal of Economics*, *12*, 445–466.

Scotchmer, S. (1985a). Profit-maximizing clubs. *Journal of Public Economics*, *27*, 25–45.

Scotchmer, S. (1985b). Two-tier pricing of shareed facilities in a free-entry equilibrium. *Rand Journal of Economics*, *16*(4), 456–472.

Sharkey, W. W. (1982). *The Theory of Natural Monopoly*. Cambridge University Press, New York.

Strotz, R. H. (1965). Urban transportation parables. In Margolis, J. (Ed.), *The Public Economy of Urban Communities*, pp. 127–169. Resources for the Future, Washington, D.C.

Varian, H. R. (1989). Price discrimination. In Schmalensee, R., and Willig, R. D. (Eds.), *Handbook of Industrial Organization*, Vol. I of *Handbooks in Economics*, chap. 10, pp. 597–654. Elsevier Science Publishing, New York.

Varian, H. R. (1992). *Microeconomic Analysis*. W. W. Norton & Co., New York.