Pergamon

S0308-5961(96)00004-3

Telecommunications Policy, Vol. 20, No. 3, pp. 219–231, 1996

0308-5961/96 $15.00 + 0.00

The political economy of congestion charges and settlements in packet networks

William H Lehr and Martin B H Weiss

This paper examines the case for usage-based pricing in the Internet by extending earlier work on congestion pricing in a single network to the case of multiple, competing carriers. A settlements problem arises in this context because of the need to allocate revenues among the carriers. The settlements and pricing problems are closely related. After deriving the optimal congestion prices, we discuss alternative settlements mechanisms and identify a number of the technical and atrategic issues that require further research before practical implementation of usage pricing in a multiple domain network is feasible. Copyright © 1996 Elsevier Science Ltd.

W H Lehr may be contacted at the Graduate School of Business, Columbia University, New York, NY 10027, USA (Tel: +1 212 854 4426; fax: +1 212 864 4857; email: wlehr@research.gsb.columbia.edu). M B H Weiss may be reached at the Telecommunications Program, Department of Information Science, University of Pittsburgh, Pittsburgh, PA 15260, USA (Tel: +1 412 624 9430; fax: +1 412 624 5231; email:mbw@lis.pitt.edu).

[1]This paper is also being published in the proceedings of the Twenty-Third Annual Telecommunications Policy Research Conference, edited by Gerald Brock and Greg Rosston and published by Lawrence Erlbaum Associates. We would like to thank Marjorie Blumenthal, Dave Clark, Jeffrey MacKie-Mason and Padamanthan
continued on page 220

Introduction[1]

The dramatic growth of Internet traffic and the expectation that ATM services will play an increasingly important role in the future of the Public Switched Telecommunications Networks (PSTN) are attracting new interest in the economics of pricing for packet-based services. Since the costs of these networks are largely fixed, optimal usage prices will differ from zero only to the extent that there are congestion costs.[2]

Our analysis extends the modelling framework presented by Mackie-Mason and Varian based on a single network domain to the case in which end-to-end network service is supplied by multiple, independent carriers who may have neither the information nor the incentive to cooperate in setting prices or preparing investment strategies that are optimal for the overall network-of-networks.[3] We show how it may be possible to set optimal congestion prices using only local information on costs and traffic. In addition, we examine the settlements problem that arises with multiple networks and discuss some of the difficulties this will present for effective implementation of congestion prices.

Congestion pricing for interconnect networks

Since most of the costs of constructing and maintaining an electronic communications network such as the telephone or Internet networks are largely fixed (or sunk), the carrier's marginal cost for handling additional traffic is close to zero. Therefore, uniform marginal cost pricing will not allow service providers to recover their costs. This has led to wide use of non-linear pricing strategies that usually take the form of multipart tariffs that include separate charges for access and usage. When carrier costs are not very sensitive to usage, then it is possible to recover the bulk of network costs in the form of a flat monthly access fee, and as long as the network's quality of service is unaffected by the

level of traffic, usage fees may be undesirable. However, if usage is free, then consumers will fail to take into account the full social costs of their traffic. These include the reduction in service quality that may be experienced by all subscribers as the network becomes more congested.

Network capacity is limited. As network congestion increases, customers may experience increased delays, higher error rates, or an increased probability that their traffic will be blocked. While the direct variable costs to the service provider may not be affected, this reduction in service quality may impose large social costs on the aggregate community of subscribers. If it turns out that it is either inexpensive enough or desirable for other reasons to install sufficient excess capacity that the network remains uncongested even with zero usage prices (ie consumer demand for bandwidth is finite at zero prices), then these social costs will be small. On the other hand, if the network is capacity-constrained, it may be desirable to charge usage prices that reflect the higher social costs associated with increasing congestion.

There are a number of solutions available for allocating scarce bandwidth among competing users. One of the most obvious is 'first come, first served'. In traditional connection-oriented telephone networks, each customer receives a fixed allocation of bandwidth until capacity is exhausted. Additional calls are blocked. While simple to implement, this strategy does not discriminate among traffic that may differ widely in its value to customers. This can lead to an inefficient allocation of bandwidth and can encourage wasteful investments by customers who must compete for the scarce bandwidth. High value uses may be driven to invest in private networks in order to guarantee access, which could result in higher costs for those who continue to rely on the public network.

A centralized call-admission or traffic-control policy could control this directly, but this would require too much information regarding the exact nature of consumer demands. One obvious alternative is to offer priority pricing: higher prices for higher quality of service and preferential access to bandwidth. This induces consumers to self-sort their traffic in order of value, which can result in significant benefits to both classes of subscribers. Another alternative is peak load or congestion pricing where users are charged prices that vary with time and the availability of resources. When capacity is scarce, prices should be higher to reflect the increased social costs of congestion. Telephone networks implement a version of this in the form of off-peak discounts for evening and weekend calling.[4]

Specifying the appropriate congestion price makes it possible to decentralize decision-making by forcing subscribers to internalize the full social costs (ie excess congestion) imposed on all subscribers to the network. Below, we show that with appropriate assumptions, it may be possible to compute these prices using only knowledge about local demand and capacity cost conditions. While the rationale for positive congestion prices is derived from the negative impact congestion may have on all users of the network-of-networks, it is not usually necessary to know individual responses to increased congestion in order to set prices. This is important, since the individual responses to congestion are not directly observable.

MacKie-Mason and Varian provide an analysis of congestion pricing in a single network. Their analysis assumes that all network costs are fixed and that subscribers benefit when they originate calls but suffer

continued from 219
Srinagesh for helpful comments and suggestions.

[2]A diverse mix of economists, engineers and computer scientists have proposed a variety of different approaches for implementing congestion-sensitive pricing in computer networks. See, for example: Bohn, Braun, H, Claffy, K and Wolff, S. 'Mitigating the coming Internet crunch: multiple service levels via precedence' technical report, UCSD, San Diego Supercomputer Center and NDF, Santiago (1993); Clark, D 'Adding service discrimination to the Internet' paper presented to Twenty-Third Annual Telecommunications Research Policy Conference, Solomons Island, MD (October 1995); Cocchi, R, Estrin, D, Shenker, S and Zhang, L 'Pricing in computer networks: motivation, formulation, and example' technical report, University of Southern California, Los Angeles (October 1992); Estrin, D and Zhang L 'Design considerations for usage accounting and feedback in internetworks' *ACM Computer Communications* 1990 **20** (5) 56–66; MacKie-Mason, J and Varian, H 'Some economics of the Internet' technical report, University of Michigan, MI (April 1993); MacKie-Mason, J and Varian, H. 'Economic FAQs about the Internet' *Journal of Economic Perspectives* 1994, **8** (3) 75–76; Parris, C and Farari, D 'A resource based pricing policy for real-time channels in a packet-switching network' technical report, International Computer Science Institute, Berkely (1992); Parris, C, Keshav, S and Ferrari, D 'A framework for the study of pricing in integrated networks' technical report TR-92-016, International Computer Science Institute, Berkeley (1992); Shenker, S, Clark, D, Estrin, D and Herzog, S 'Pricing in computer networks: reshaping the agenda' paper presented to Twenty-Third Annual Telecommunications Research Policy Conference, Solomons Island, MD (October 1995)

[3]See MacKie-Mason, J and Varian, H 'Pricing congestible network resources *IEEE Journal on Selected Areas in Communications* 1995 **13** (7) 1141–1149. Hereafter this will be referred to simply as MacKie-Mason and Varian in the text, unless otherwise noted.

[4]When traffic patterns are relatively predictable, peak load prices, such as those used in telephony, are possible. When the congestion is unpredictable, dynamic prices may be necessary.

when network congestion increases. Congestion increases with network utilization, measured as the ratio of aggregate traffic to network capacity. Since the only beneficiary of an additional call is the originator, and since each additional call increases network congestion, the social externality is unambiguously negative, which provides the justification for positive congestion prices.[5] In their framework it is relatively straightforward to demonstrate that the efficient uniform congestion price is a function of aggregate demand, total capacity costs and network capacity. It is not necessary to observe individual consumer demands in order to set optimal congestion prices for an efficiently sized network. Since the individual demands are not readily observable by the service provider, this result is important. Although it is unclear how the carrier selects the efficiently sized network, it is plausible that the carrier might be able to forecast aggregate demand for a single network domain.

We extend the analysis of MacKie-Mason and Varian to the case of M network domains, which raises several important issues. First, once there are two or more networks, it is no longer clear how one should measure the congestion experienced by a subscriber. In principle, we might expect it to vary depending on the type of calls made (ie on-net or internet), the route followed by the call and the capacities of the various subnetworks.[6] Second, there is the additional problem of settlements, or determining how usage, and potentially access revenues, should be distributed among the multiple carriers. In a dynamically stable long-run equilibrium, each must recover sufficient revenues to cover its network costs. In general, this will require transferring revenue among the carriers. The mechanism chosen for mediating these transfers (eg on the basis of calls handled) may affect carriers' incentives to manipulate their congestion status, which in turn may influence the setting of congestion prices. To address these issues, we modify the earlier modelling framework as follows.

Let there be M networks, each of which has N_i total subscribers. A type 'ij' subscribers makes calls that originate on network 'i' and terminate on network 'j'. These calls are transported across each of the networks along the route followed by type 'ij' calls. Let $R(ij) \subset M$ denote the subset of networks that are included in the route of call 'ij'. To simplify the analysis we assume each subscriber makes a unique type of call and that the call follows a unique path through the network-of-networks.[7] Let $Z = \{ij \text{ such that } i,j \in M\}$ designate the set of all possible types of calls. The total number of subscribers on the ith network is given by $N_t = \Sigma_{j \in M} N_{ij}$. Let $U^{ij} = \mathrm{U}^{ij}(\mathrm{x_{ij}}, Q^{ij})$ be the utulity of a type 'ij' consumer, where x_{ij} is the number of type 'ij' calls and Q^{ij} is the congestion experienced by type 'ij' calls. Following MacKie-Mason and Varian, assume that utility is weakly increasing in calls originated and is weakly decreasing in the level of congestion (ie $\partial U^{ij}/\partial x_{ij}\ 0 \geqslant 0$ and $\partial U^{ij}/\partial Q^{ij} < 0$).[8]

The level of congestion, Q^{ij}, provides an inverse proxy for the quality of service experienced by 'ij' calls. It could be measured in a wide variety of ways such as the level of average delay, the maximum potential delay, the bit error rate, the delay jitter, the blocking probability, or some weighted average of all of these. In general, we might expect it to be a weakly increasing function of the volume of each type of traffic and a weakly decreasing function of each network's capacity. We further specialize the analysis by assuming that congestion

[5]If the recipients of calls also benefit and this benefit is sufficiently large, then the social externality from additional calls may be positive. Sringagesh notes that this is one of the rationales for zero settlements among Internet service providers. See Srinagesh, P 'Internet cost structures and interconnection arrangements' in Brock, G (ed) *Toward a Competitive Telecommunications Industry: Selected Papers from the 1994 Telecommunications Policy Research Conference* Lawrence Erlbaum Associates, Hillsdale, NH (1995)

[6]We use 'internet' (uncapitalized) to refer to communications across semi-autonomous network domains. The Internet is the worldwide TCP/IP packet-switched collection of networks that have evolved from the research-based Department of Defence-funded ARPANET. The Internet is just the best known of the many potential internets to which our analysis may apply.

[7]The assumption that each subscriber makes a single type of call is less restrictive than it may at first appear, since a 'real world' subscriber who makes multiple types of calls may be modeled as several different subscribers as long as he or she does not regard different types of calls as close substitutes. This seems reasonable for most types of calling (ie a caller in New York does not regard calls to California and Florida as substitutes). The assumption of unique routing may be extended to include connectionless traffic if time intervals are suitably short and $R(ij)$ is allowed to change over time.

[8]We will assume that subscribers ignore the effect their traffic has an overall congestion since N_i and perhaps N_{ij} are large [or, $(\partial U^{ij}/\partial Q^{ij})(\partial Q^{ij}/\partial x_{ij})$ is close to zero]. Note that this does not imply that the aggregate effect on all subscribers of additional congestion is small.

is measured in terms of the average end-to-end delay and that this is simply the sum of the average delay expected at each switching node along the call's route, or,

$$Q^{ij} = \sum_{k \epsilon R(ij)} D[Y_k], \tag{1}$$

where $D[Y_k]$ is the average delay on the kth subnetwork along the route. We assume that $D[.]$ is a continuous, monotonically increasing function of network utilization, which is defined as the aggregate traffic handled by network 'k' divided by its capacity (ie $Y_k = X_k/K_k$).

The aggregate traffic carried by the ith network, X_j, consists of on-net and internet traffic. On-net traffic both originates and terminates on the same network. The internet traffic may be divided into traffic that originates (terminates) on the ith network, but terminates (originates) on another network and pure transit traffic. The total traffic that originates on network 'i' equals $X_i^{On} + X_i^{Off}$, where $X_i^{On} = N_{ii}x_i$ is the *on-net* traffic and $X_i^{Off} = \Sigma_{k \epsilon M, i \neq k} N_{ik}x_{ik}$ is the internet traffic. The internet traffic that either terminates on network 'i' or is pure transit traffic is given by $X_i^{In} = \Sigma_{kj \epsilon z, k \neq i, i \epsilon R(kj)} N_{kj}x_{kj}$. Therefore $X_i = X_i^{On} + X_i^{Off} + X_i^{In}$.

Assume two-part tariffs and voluntary participation and that the 'sender-pays', so that the surplus realised by consumer 'ij' is $U^{ij}(x_{ij},Q^{ij}) - p_{ij}x_{ij} - T_i \geq 0$ in equilibrium, where p_{ij} is the total congestion charge for call 'ij' and T_i is the fixed access charge for network 'i'.

Assume that all network costs are fixed and that the costs of each subnetwork depend only on the capacity of that subnetwork. Let the cost of the ith network be described by a continuous, differentiable function $C^i(K_i)$.[9]

Finally, we define social welfare as the sum of consumer and producer surplus and assume that there are no exernal subsidies allowed.

With the above assumptions and in the absence of settlements, the profit realised by the ith network service provider can be computed as the sum of access and usage revenues less network costs:

$$\Pi^i = N_i T_i + \sum_{J \epsilon M} N_{ij} p_{ij} x_{ij} - C^i(K_i). \tag{2}$$

The third term is the net lump sum transfer received by the ith network, i. The assumption of voluntary participation implies that Π^i must be weakly positive in equilibrium. Total welfare may be computed as:

$$W = \sum_{i \epsilon M} \sum_{j \epsilon M} N_{ij}(U^{ij} - p_{ij}x_{ij} - T_i) + \sum_{i \epsilon M} \Pi^i. \tag{3}$$

In the absence of settlements, one finds the optimal congestion prices for an equilibrium-sized network from inspection of the first order condition for maximizing social welfare with respect to each type of traffic. Each of these first order conditions is of the form:

$$\frac{\partial W}{\partial x_{ij}} = 0 = N_{ij}\frac{\partial U^{ij}}{\partial x_{ij}} + \sum_{\substack{lk \epsilon Z \\ lk \neq ij}} N_{lk}\frac{\partial U^{lk}}{\partial Q^{lk}}\frac{\partial Q^{lk}}{\partial x_{ij}}. \tag{4}$$

The second term is the negative externality imposed on other network subscribers from increased congestion when type 'ij' consumers increase

[9] In a more general model, we might not expect network costs to be separable as assumed here. Also, we might expect more complex interactions among different types of traffic and capacity in the determination of call-specific congestion. Furthermore, computing the least cost route for a call may be quite difficult, since it amounts to optimally routing traffic so as to minimize congestion costs.

their calling. In order to induce a type '*ij*' subscriber to internalize the effects of her calling, congestion prices should be set so that:

$$P_{ij}^* = -\sum_{\substack{lk \epsilon Z \\ lk \neq ij}} \left(N_{lk} \frac{\partial U^{lk}}{\partial Q^{lk}} \frac{\partial Q^{lk}}{\partial x_{ij}} \right) - (N_{ij} - 1) \frac{\partial U^{ij}}{\partial Q^{ij}} \frac{\partial Q^{ij}}{\partial x_{ij}}. \quad (5)$$

The first term on the right side of Equation (5) represents the congestion externality imposed on other subscribers whose traffic is carried on the *i*th network, while the later term is the congestion externality imposed on other type '*ij*' subscribers. Substituting further for Q^{ij} in (5) and rearranging yields:

$$P_{ij}^* = -\sum_{n \epsilon R(ij)} \frac{D^n{}_Y}{K_n} \left(\sum_{\substack{lk \epsilon Z \\ n \epsilon R(lk)}} N_{lk} U^{lk}{}_Q \right), \quad (6)$$

where $D^n{}_y = \partial D(Y_n)/\partial Y_n$ and $U^{lk}{}_Q = \partial U^{lk}/\partial Q^{lk}$. Note that, since network utilization may vary, we cannot assume that the marginal increase in delay is constant for all networks. Therefore, we retain the '*n*' superscript to remind ourselves that D_y ought to be computed for each network along the route of call '*ij*'. If we further assume that network service providers earn zero profits, ie that the markets are contestable,[10] then we can compute the optimal access charge incorporating the optimal values for X, p and K into the service providers' profit functions.[11]

With a single network as in MacKie-Mason and Varian, the optimal congestion price is given by:

$$p^* = -\frac{N-1}{K} \frac{\partial U}{\partial Q} \frac{\partial D}{\partial Y}. \quad (7)$$

In the case where $M = 2$, there are only four types of calls: '11' and '22' on-net traffic; and '12' and '21' internet traffic. We can use the formula in Equation (7) to compute the optimal congestion prices for the three types of traffic as follows:

$$P_{11}^* = -\frac{D^1{}_Y}{K_1} (N_{11} U^{11}{}_Q + N_{12} U^{12}{}_Q + N_{21} U^{21}{}_Q) \quad (8)$$

$$P_{22}^* = -\frac{D^2{}_Y}{K_2} (N_{22} U^{22}{}_Q + N_{12} U^{12}{}_Q + N_{21} U^{21}{}_Q). \quad (9)$$

$$P_{12}^* = -\frac{D^1{}_Y}{K_1} (N_{11} U^{11}{}_Q + N_{12} U^{12}{}_Q + N_{21} U^{21}{}_Q)$$

$$-\frac{D^2{}_Y}{K_2} (N_{22} U^{22}{}_Q + N_{12} U^{12}{}_Q + N_{21} U^{21}{}_Q) \quad (10)$$

$$= P_{21}^* = P_{12}^* + P_{22}^*$$

Thus, the optimal congestion price for internet calls should be equal to

[10] See Baumol, W, Panzar, J and Willig, B *Contestable Markets and the Theory of Industry Structure* Harcourt, Brace and Jovanovich, New York (1982)

[11] One must check that each consumer's surplus is weakly positive such that participation is not an issue. We assume that this is the case.

the sum of the congestion prices for on-net calls. This is intuitively satisfying because an internet call congests both networks, whereas an on-net call congests only the network that carries it. This result generalizes to the case of M networks: to find the optimal congestion price for a call 'ij', one should add the optimal on-net congestion prices for each node along the route [ie for the subset of networks in $R(ij)$].

When the above pricing results are combined with the first order conditions used to compute the welfare maximizing levels of capacity for each of the M networks, we obtain the following relationship:

$$\frac{\partial W}{\partial K_j} = 0 = \sum_{\substack{lk \epsilon Z \\ j \epsilon R(lk)}} \left(N_{lk} U^{lk}{}_Q \frac{\partial Q^{lk}}{\partial K_j} \right) - \frac{\partial C^j(K_j)}{\partial K_j}$$

$$= - \frac{X_j D^j{}_Y}{(K_j)^2} \left(\sum_{\substack{lk \epsilon Z \\ j \epsilon R(lk)}} N_{lk} U^{lk}{}_Q \right) - \frac{\partial C^j(K_j)}{\partial K_j} \qquad (11)$$

or,

$$p_{ii}{}^* = \frac{\partial C^i(K_i)}{\partial K_i} \frac{K_i}{X_i} \cdot \qquad (12)$$

This is analogous to the result in MacKie-Mason and Varian and shows that it is possible to compute the optimal on-net congestion charge based on local information (ie without direct knowledge of the utility functions for the individual subscribers) at equilibrium. As long as each subnetwork charges each packet it carries $p_{ii}{}^*$, the total congestion revenues collected by network 'i' will provide it with the proper signal for when to expand capacity (ie when congestion revenues exceed the value of the subnetwork's capacity valued at the marginal cost of additional capacity).

Three points are worth noting about this result. First, the optimal solution requires that internet traffic should face higher end-to-end congestion charges because it results in more congestion per minute than does on-net traffic. In general, each type of traffic that has a different impact on overall congestion should face a different end-to-end congestion price. This is a form of 'congestion priority pricing', which is analogous to other priority pricing schemes in its intent but is motivated by a slightly different need. In priority pricing, subscribers who are less congestion-sensitive accept a lower quality of service in return for a lower price. In the example cited above, it would be optimal to charge different rates for internet and on-net traffic even if all consumers had identical preferences with respect to congestion.

Second, the sub-networks will need to account for all of the traffic that passes across their networks in order to set efficient local congestion prices, and subscribers will have to be billed for the sum of these prices along the least cost route. One solution is to have a 'pay-as-you-go' billing scheme, where each network charges each packet handled its on-net congestion price and bills the consumer directly. Alternatively, the customer could be billed by the originating network, but then the orginating network would need to know what the sum of the congestion prices is along the rest of least cost route (ie $p_{ij}{}^* - p_{ii}{}^*$) in order to set the appropriate price for a type 'ij' call.

If there are at most two networks involved in every internet call (ie there are no transit networks), networks could bill each other for terminating calls.[12] This would provide each subnetwork with the information about the appropriate termination charge for a call, and the total congestion revenue collected would provide an accurate signal of whether it was advisable to expand capacity.

Another solution is to have the networks continuously update each other regarding their congestion charges, which would allow the originating network to compute p_{ij}^* directly. This may be the case in a least cost routing environment. If routing is hop-by-hop, then the appropriate congestion charge could be passed back up the chain if each node billed traffic the sum of its on-net cost plus the cost charged to terminate the call at the next link in the chain. For example, in a call that will be routed from 1 to 2 to 3, network 2 should charge network 1 the price $p_{22}^* + p_{33}^*$, which will allow network 1 to compute the appropriate end-to-end charge without direct knowledge of network 3's congestion status.

In all of these solutions, it is possible for the networks to exchange the required information in the form of traffic accounting data without actually making what might amount to sizable revenue transfers in both directions. However, it is important for the networks to account for the congestion charges associated with terminating or transmitting traffic that originates on other networks. Failure to include this traffic may either result in on-net prices that are too high or the failure to invest in adequate network capacity when such investment is appropriate.

Third and finally, while the ability to compute optimal prices based solely on local conditions holds at equilibrium, it is not clear how equilibrium would be attained in a network-of-networks without the sharing of aggregate demand information among the carriers. Although MacKie-Mason and Varian do not address this point directly, it seems somewhat more plausible in the context of a single network domain that the carrier would be able to forecast aggregate demand. In the network-of-networks context, the individual carrier would need to forecast the demands of all subscribers on all networks in order to identify the efficient configuration of subnetwork capacities. While a better understanding of how this equilibrium solution might emerge, and its stability properties is obviously important if congestion pricing is to prove useful, further consideration of pricing dynamics is beyond the scope of the present paper. The result presented here is most useful in highlighting the additional complexities introduced when network ownership is fragmented.

Optimal congestion prices and settlements

To understand why a settlements problem arises in a network-of-networks, it is sufficient to consider a very simple example with just two networks. Assuming no settlements, optimal congestion prices and origination-network billing, each network will earn profits of:

$$\Pi^1 = N_1 T_1 + X_1^{on} P_{11}^* + X_1^{off}(p_{11}^* + p_{22}^*) - C^1(K_1) \quad (13)$$

$$\Pi^2 = N_2 T_2 + X_2^{on} P_{22}^* + X_2^{off}(p_{11}^* + p_{22}^*) - C^2(K_2). \quad (14)$$

If the network-of-networks is to recover its costs without external subsidies, then the sum of the profits of the constituent networks must

[12]With three networks, the pure transit network could bill the customer and then pay the originating and terminating congestion charges. A version of this occurs in long-distance telephone when the long-distance company pays the originating and terminating local exchange carriers a per minute access charge.

be weakly positive. In the absence of settlements, the profits of *each* network must be weakly positive. This imposes a stronger constraint on the optimization problem and may require distorting the optimal solution in order to be satisfied.

If the markets were contestable (free-entry) or under appropriate rate of return regulation, service providers might be expected to earn zero economic profits. Setting $\Pi^1 = 0$, substituting for the efficient congestion prices and rearranging yields the following result (which is analogous to the result in MacKie-Mason and Varian:

$$\frac{T_1 N_1}{C^1(K_1)} = 1 - \frac{\partial C^1(K_1)}{\partial K_1}\frac{K_1}{C^1(K_1)} + \frac{p_{11}{}^* X_2{}^{off} - p_{22}{}^* X_1{}^{off}}{C^1(K_1)}. \qquad (15)$$

The left hand side gives the share of network costs that must be recovered via the flat access fees in order for the network to recover its costs. The second term on the right drops out if there is only one network, or if traffic flows are balanced and the optimal congestion prices are identical. In either of these special cases, the share of network costs that are recovered via the flat access fee increases towards one as the ratio of marginal to average capacity costs goes to zero. In the multiple network case, however, it is unlikely that traffic flows would be identically balanced or that the optimal congestion prices will be equal.

In the fully symmetric case with equal numbers of on-net and internet callers and identical costs for each network, the optimal congestion prices, access fees, traffic and capacity for each network will be identical. There will not be a settlements problem. Consider what happens, however, if the subscribers are distributed asymmetrically such that a larger share of the internet callers is located on network 1. Under our assumptions, the network congestion caused by a call depends on the route followed but not the direction of the route (ie call '12' causes the same congestion as call '21'), so this change should not affect the optimal access and congestion charges faced by consumers.[13] Under the original solution, however, network 2 will fail to recover its costs.

In the absence of settlements, there are a number of approaches that may be used to resolve this problem. First, if participation is not an issue, we could allow asymmetric access charges, with network 2 charging an access fee that is sufficient to recover its higher costs.[14] While this solution may be efficient, it may not be perceived as equitable. One could argue that it is unfair that consumers on network 2 face higher access charges, since consumers on network 1 also benefit from the reduction in overall congestion when network 2's capacity expands.

Second, if we constrain ourselves to uniform access pricing, it may still be possible to implement the efficient capacity and congestion pricing solution by charging higher access fees to all subscribers. In this case, we would need to prevent entry competition for network 1, since it will earn positive profits at p^* and the new, higher T^{**}.

Third and finally, if we constrain ourselves both to free-entry and to uniform pricing, then it will be optimal generally to modify both usage *and* access fees, and in general we will not be able to achieve the same level of total surplus as in the unconstrained problem. This problem arises because in a zero-profit equilibrium it is possible that sizable congestion revenues will be collected from subscribers in order to

[13]We are assuming here that the level of network capacity costs depends on traffic patterns and not on the number of subscribers. Although in general we might expect network costs to depend both on the number of subscribers and the capacity, K_j (which itself may depend on the number of subscribers), this need not be the case for several reasons. First, K_j refers to the capacity that is relevant for determining the level of network congestion. This capacity might be the size of the switch, which may depend on X and not the number of subscribers that generate X. Second, capacity may have to be added in fixed increments, and so equal capacity may be optimal for differing numbers of subscribers over a relatively large range. Third and finally, all traffic may be internet traffic, in which case both networks need identical congestion capacity, because all calls transit both networks.

[14]If consumers could move freely, then we would end up with the fully symmetric case. However, subscribers may not be freely mobile.

induce them to properly internalize the welfare implications of increased calling. These congestion revenues will permit firms to charge lower access fees than would be necessary in the absence of congestion charges, but the sum of these congestion charges and access fees may be insufficient to recover the costs of all of the networks in the optimal solution. In a 'sender-keep-all, no settlements' world it would be possible for an uncongested, upstream network that originates a disproportionate amount of traffic to collect most of the congestion revenue.

Implementation issues

The discussion in the preceding two sections demonstrated that congestion pricing in a network-of-networks is significantly more challenging than may have been apparent from consideration of the case of a single network domain. In the following two sections, we identify additional complications that will need to be addressed before it is practical to implement congestion pricing. Broadly, these can be classified as technical and strategic. Our goal is to suggest important topics for further research, rather than to posit solutions, which in any case is well beyond the scope of the present paper.

Technical implementation considerations

The result that decentralized congestion pricing is optimal is important from a practical perspective. It means that the decentralization of network control, by itself, does not necessitate the sharing of information of the congestion of neighboring networks. At the optimum, each network can compute a single congestion price based on local demand and cost information. While this result is encouraging, there are numerous other practical problems that would need to be addressed.[15] A partial list includes the interaction among applications types, network architecture and accounting, type of service considerations and accounting overhead (ie how much it will cost to modify network hardware and software). These concerns (and others) have given rise to arguments that simple packet counting is not an adequate basis for settlements.[16]

In addition to these issues, there are several other considerations that require further investigation.

- Congestion prices work by forcing subscribers to internalize the congestion externality caused by their use of the network. If the total congestion price of a packet is the sum of the congestion prices of the networks it traverses, the user must be aware of the congestion price before the packet is sent. This requires that all price information be continuously available to all users (or subnetworks to which users are attached) *and* that the user (or subnetwork) know the route a packet will take in advance. The first requirement places an information flow requirement on all of the networks that may be substantial, depending on how the congestion pricing scheme is implemented. The second requirement is reasonable for connection-oriented network services but may not be for connectionless network services, depending on the routing scheme used and the frequency with which congestion prices change.
- Even if congestion prices are implemented, and price information is dispersed appropriately, there is still the question of billing for network service. There have been a number of approaches that have

[15]Estrin and Zhang *op cit* Ref 2 have considered some of these.

[16]See, for instance remarks attributed to Vinton Cerf in Cook, G 'Summary of the September 1995 COOK report' distributed on the *telecomreg* newsgroup, September 3 1995. This report also raises the issue of different 'business models' of the internet service providers, arguing that MCI's Internet network, as a predominant 'transit' network, is currently unprofitable, raising the pressure for some sort of settlements scheme.

been proposed for accounting and billing in networked information systems.[17] Before any of these approaches can be applied, however, an overall collection and billing strategy must be identified.

- Computing $\partial C^j(K_j)/\partial K_j$ is likely to be difficult in a complex subnetwork consisting of many components. While we use the term 'capacity' fairly loosely here, its precise definition is more elusive, since 'capacity' can be affected by network management, congestion control techniques, etc in addition to direct investments in network facilities.
- Our solution does not easily adapt to multicast.
- We assume that any 'receiver-pays' scheme will be handled externally, perhaps using technology like NetBill.[18]

The way in which these details are resolved matters. If the originating network supplies the end-to-end price to the user and performs the billing, settlements may be necessary. If each individual network announces price and bills separately, then additional user software is necessary to present a consolidated congestion price (and perhaps a bill) to the end-user.[19]

The congestion pricing we have analyzed here does not include multiple service classes, such as 'real time' or 'best effort'. It is widely anticipated by computer science researchers that some form of performance guarantee will be needed to implement real time traffic.[20] Parris and Ferrari argued that different service classes require different prices.[21] Stahl and Whinston have considered client-server computing with priority classes. The structure of their analysis can inform the problem of multiple service classes in networks with congestion externalities as well.[22]

Strategic implementation considerations

In the preceding discussion, we have assumed that network providers do not have market power and hence will not be able to bias their pricing, network capacity or interconnection decisions either to extract consumer surplus or to protect surplus profits. If market power is significant in a privatized Internet, then there will be myriad ways in which service providers may seek to distort either congestion pricing or the settlements mechanism. For example, a transit network that controlled a bottleneck facility would have an incentive to distort its prices for access (interconnection) and usage fees in order to extract monopoly rents. It may charge lower or higher than optimal usage fees, depending on the relationship between inframarginal and marginal subscriber responses.

If monopoly rents are collected by any of the carriers, then the settlements mechanism would provide a vehicle for distributing those rents. Bargaining over the distribution of these rents is likely to prove contentious, which will further complicate implementation of a settlements process. Introducing settlements into network profit calculations will influence their behavior. From the discussion in the preceding section, it should be clear that monitoring individual subscriber or subnetwork behavior would be difficult, and hence carriers may have an incentive to misrepresent their traffic/congestion status in order to capture a larger share of any settlements revenue. There is a principal-agent problem that must be resolved. Failure to agree on an appropriate settlements mechanism may cause the network-of-networks to fragment.

[17]See, for instance: Edell, R, McKeown, N, Variaya, P 'Billing users and pricing for TCP' *IEEE Journal on Selected Areas in Communications* 1995 **13** (17) 1163–1175; Mills, C, Hirsh, D, and Ruth, G 'Internet accounting: background' technical report RFC 1272, Network Working Group (1991); Ruth, G and Mills, C 'Usage-based cost recovery in internetworks' *Business Communications Review* 1995 **XX** (July 1992) 38–42; Sirbu, M and Tygar, 'Netbill: an internet commerce system optimized for network delivered services' paper presented to Workshop in Internet Economics, Massachusetts Institute of Technology, Boston (Summer 1995)

[18]See Sibru and Tygar *op cit* Ref 17

[19]This is, in effect, how the telephone network presently works. Users pay a fixed network access fee directly to the local telephone operating company and receive a separate statement (often in a consolidated bill) from the interexchange carrier. This bill includes all settlements between the carriers. See, for instance, Danielsen, K and Weiss, M 'User control modes and IP allocation' technical report, University of Pittsburgh, Pittsburgh (March 1995).

[20]See, for example, Ferrari, D 'Real-time communication in an internetwork' *Journal of High Speed Networks* 1992 **1** (1) 79–103; Field, B 'A network channel abstraction to support application real-time performance guarantees' PhD thesis, University of Pittsburgh, Department of Computer Science, Pittsburgh (1994)

[21]Parris and Ferrari *op cit* Ref 2

[22]Stahl, D and Whinston A 'An economic approach to client-server computing with priority classes' technical report, University of Texas at Austin (1992)

In the past, concerns over excess market power provided the justification for regulation of the cable television and telephone industries. In recent years, disaffection with traditional regulatory remedies and advances in technology that have reduced entry barriers have encouraged a trend towards increased reliance on market forces. While the difficulties posed by imperfect competition are worthy of significant research attention, they go beyond the scope of the present paper. However, even if we restrict ourselves to the (perhaps dubious) case of contestable carrier markets, we cannot presume that all subscribers will be equally represented or influential in determining how future networks will evolve.

For example, in our model, there is a fundamental tension between subscribers who make different types of calls. On-net and internet callers each would like to see the other's traffic minimized and hence would prefer to see the other face higher prices. This may have implications for customer attitudes towards the efficient implementation of congestion pricing and towards the debate about emerging notions of 'universal service' for the Internet.[23] As noted above, efficient prices should discriminate among on-net and internet traffic, and non-zero settlements offer one mechanism for implementing these higher prices.

Let us suppose that the network community can be convinced of the advisability of congestion pricing and that the debate has turned to the need to discriminate among different types of traffic.[24] Since efficient congestion pricing implies that internet traffic should face higher prices, these callers would have an incentive to argue against price discrimination, while on-net subscribers would take the opposite position. Since the settlements mechanism that is chosen is likely to affect the feasibility of implementing price discrimination, there may be a bias from 'internet-type' callers in favor of zero-settlements mechanisms.[25] Consider what might happen in negotiations between the subscriber communities of a large and a small network with symmetric calling among pairs of subscribers. In aggregate, subscribers on the larger network are more likely to make on-net calls, while subscribers on the smaller network are more likely to make internet calls. Thus, under congestion pricing, subscribers on the larger network should press for a complex settlements mechanism that facilitates charging for termination traffic, while subscribers on the smaller network may argue for zero settlements. The point of this discussion is to suggest how, even in the absence of market power by service providers, the political debate over optimal pricing may be distorted by private economic interests.

The failure to adopt optimal congestion prices may influence the choice of where subscribers choose to originate their traffic, although not all subscribers are likely to face the same flexibility. For example, optimal congestion prices should be identical regardless of the direction in which a particular calling route is followed. If, however, $p_{ij} > p_{ji}$, then sophisticated callers will have an incentive to originate their calls from network j. It is not necessary for a caller to physically locate on another network, since he or she could use an inexpensive call to set up the return origination call.[26] Generally, rate arbitrage that results in similar end-to-end congestion charges for traffic with similar congestion (quality-of-service) characteristics would be welfare-improving. However, such arbitrage may not occur on a sufficiently large scale and may leave unsophisticated subscribers at a disadvantage.

Content providers are another class of sophisticated subscribers who

[23]There is sizeable community of Internet users that oppose usage-based pricing. Many of these users are concerned about the effects of usage-based pricing on the modes of behavior (such a mailing lists) that they perceive to be valuable. See Love, J 'Future internet pricing' available via gopher:// essential.essential.org: 70/ 0R0-12615--/pub/listserv/tap-info/950310, 10 March 1995

[24]We ignore the accounting and implementation costs associated with usage pricing. These may be substantial and when included in the cost/benefit analysis may make usage pricing inefficient. Assessing the magnitude of these costs is clearly an important area for further research.

[25]This bias may be partially (or wholly) offset if the uniform on-net price or access fees rise in order for the network to recover its total costs.

[26]A number of entrepreneurs offered such services to international callers to arbitrage international telephone settlements that resulted in higher prices for calls that originated internationally.

may seek to influence the setting of usage prices.[27] Generically, we might presume that they would like to see relatively low network access and usage fees so that consumers have more surplus to spend on content. Ideally, they might like to see network services provided free (subsidized by general tax revenues that would include non-subscribers). Alternatively, if the typical content customer is an infra-marginal consumer of network services, they may prefer higher than optimal access fees in return for lower than optimal usage fees. Although this scenario need not be the case, we suggest it to illustrate why the establishment of usage pricing is likely to be contentious.

Summary and conclusions

We believe usage pricing is both desirable and unavoidable for the Internet. We also believe that there is still much research that needs to be done to better understand both the theoretical and practical issues that arise in a network-of-networks. This paper offers a first step towards examining the dual problem of congestion pricing and settlements in such an environment. We proceed by extending the analysis of a single network domain included in MacKie-Mason and Varian to the case of multiple networks. This analysis shows that the end-to-end congestion price should equal the sum of the on-net congestion prices of each of the networks along the route. In an efficiently sized network, these prices may be computed using only local cost and traffic information. This is important if network control is to be decentralized.

In the absence of centralized coordination, the networks need to share congestion pricing information so that the originating networks can know what price to set for end-to-end service. A settlements process that requires networks to bill each other for terminating traffic offers one mechanism for conveying this information. This provides one rationale for the linkage between the two problems. A second rationale stems from the need for each network to recover its costs. If prices are set so as to induce optimal consumer behavior by forcing them to internalize the welfare implications of their behavior for the network-of-networks, then individual firms may fail to recover sufficient revenue in the absence of settlements.

Even in a world where firms do not have market power, revenue transfers among service providers (ie settlements) are likely to be necessary, and since the amount of revenue transferred is likely to depend on both the volume of traffic and the price faced by consumers, congestion pricing and settlements issues are not readily separable. We demonstrate this using a simple case of two networks. Further, we argue that the nature of the settlements problem depends on the technology of the networks being used to deliver service as well as the design of the settlements mechanism.

Our analysis focused on the case where carriers do not have market power. If this assumption is not valid, then the problem becomes considerably more complex, since strategic interactions among the service providers must be considered. In addition, we concentrated on the situation where a network provides a single type of service, as in today's Internet. If multiple service classes exist, as may be necessary with the emerging ATM-based networks, or if a scheme such as the 'smart-market' or 'precedence' is used to provide price-based priority, additional factors may need to be considered.[28] Finally, our analysis is

[27]See, for example, MacKie-Mason, J and Varian, H 'Network architecture and content provision: an economic analysis' paper presented to Twenty-Third Annual Telecommunications Research Policy Conference, Solomons Island, MD (October 1995).

[28]For smart-markets see MacKie-Mason and Varian *op cit* Ref 27 and for precedence pricing see Bohn *et al op cit* Ref 2

static and does not consider the important question of how the efficient pricing equilibrium is attained nor whether it is stable. A dynamic analysis raises numerous technical problems that must be solved (not the least of which is the user interface). There are many new economic issues that arise in a dynamic analysis, particularly if a settlements strategy is included explicitly in the analysis. There is clearly much more work that needs to be done in the area of generalizing this analysis from an economic perspective and in applying it to specific network implementations, both statically and dynamically.