Pergamon

0308-5961(96)00002-X

Telecommunications Policy, Vol. 20, No. 3, pp. 183–201, 1996

0308-5961/96 $15.00 + 0.00

Pricing in computer networks: reshaping the research agenda

Scott Shenker, David Clark, Deborah Estrin and Shai Herzog

As the Internet makes the transition from research testbed to commercial enterprise, the topic of pricing in computer networks has suddenly attracted great attention. Much of the discussion in the network design community and the popular press centers on the usage-based versus flat pricing debate. The more academic literature has largely focused on devising optimal pricing policies; achieving optimal welfare requires charging marginal congestion costs for usage. We contend that the research agenda on pricing in computer networks should shift away from the optimality paradigm and focus more on structural and architectural issues.

S Shenker may be contacted at Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304-1314, USA (Tel: +1 415 812 4840; fax: +1 415 812 4471; email: shenker@parc.xerox.com). D Clark may be contacted at the Laboratory for Computer Science, Massachusetts Institute of Technology, 545 Technology Square, NE43-507, Cambridge, MA 02139, USA (Tel: +1 617 253 6003; fax: +1 617 253 2673; email: ddc@lcs.mit.edu). D Estrin and S Herzog may be contacted at the University of Southern California, ISI, Los Angeles, CA, USA (email: estrin@usc.edu/herzog@isi.edu).

We are grateful to Jeffrey MacKie-Mason, Hal Varian, Majory Blumenthal, Padmanabhan Srinagesh and Steve Deering for insightful comments on an earlier draft. We would also like to thank Van Jacobson for early conversations on edge pricing.

continued on page 184

Introduction[1]

In a few short years, the Internet has made a dramatic transformation from nerdy enigma to trendy hangout. With its millions of users and diverse application offerings, the Internet is now seen by many pundits as the archetype of the future global information infrastructure. Because of its heavily subsidized origins, commercialization has come late to the Internet. As the Internet confronts this belated and somewhat awkward transition from research testbed to commercial enterprise, there has been much recent discussion about the role of pricing in computer networks. Numerous workshops and conferences have been held on the topic in both the academic community and the network design community; the popular press has also seized upon the issue as one of broad interest.[2]

In the popular press and in the network design community, the agenda has been dominated by debates over whether to move from the present system of charges based on the speed of the access line (so-called 'flat pricing') to basing charges on actual usage. Some contend that usage-based pricing is unnecessary and would have disastrous consequences for the Internet. Others argue that moving away from flat pricing towards usage-based pricing is essential for the Internet's efficiency and therefore is the key to its future economic viability. Unfortunately, little has been clarified by this heated debate except the depth of the participants' convictions. We hope to demonstrate in this paper that usage-based charging and flat pricing are really two ends of a single continuum, so the difference between them is not one of fundamental principle but merely of degree, and hybrids of the two approaches will likely be commonly used in the future.

The academic discussion of pricing in computer networks has concentrated on a rather different issue. This literature typically assumes the necessity of usage-based pricing and focuses on achieving optimal efficiency—maximal welfare—in certain simplified models using usage-based pricing schemes. The satisfaction a network user derives from

his/her network access depends on the nature of the application being used and the quality of service received from the network (in terms of bandwidth, delay, packet drops, etc); the network's resources are used most efficiently if they maximize the total user satisfaction of the user community. To achieve optimal efficiency, usage-based charges must equal the marginal cost of usage. Since the physical transmission of packets is essentially free, the marginal usage cost is almost exclusively a congestion cost; congestion costs are the performance penalties that one user's traffic imposes on the others. This *optimality* paradigm dominates the research agenda; much of the literature discusses pricing schemes based on computations of these marginal congestion costs.

The main purpose of this paper is to advocate shifting the research agenda away from the reigning optimality paradigm and towards a more architectural focus. We will use the phrase *pricing architecture* to refer to those components of the pricing scheme that are independent of the particular local pricing decisions and reflect nonlocal concerns, such as how receivers rather than senders can be charged for usage and how to appropriately charge multicast transmissions. These architectural issues, rather than the detailed calculation of marginal congestion costs, should form the core of the research agenda. To motivate this shift in research emphasis, we discuss both the economic issues and also the mechanistic design issues central to computer network pricing. Our treatment of these issues is designed to be accessible to both the network design community and the economic community, with the intention of providing some common context for these two communities and thereby increasing the opportunity for dialogue.

Our paper has three distinct parts. The first part critiques the optimality paradigm.[3] We first contend that usage charges may, and perhaps should, exceed marginal congestion costs. Moreover, we argue that these marginal costs are inherently accessible, and so the quixotic pursuit of their precise computation should not dominate the research agenda. The second part presents a rather different paradigm for network pricing: *edge* pricing. This term refers to where the charges are assessed rather than their form (eg usaged-based or not) or their relationship to congestion (eg marginal congestion costs or not). This emphasis reflects our belief that architectural issues are more important than the detailed nature of the charges themselves. The third portion of the paper describes two fundamental architectural issues and some preliminary design approaches. We conclude in the final section with a brief summary. Because much of our discussion requires some familiarity with network mechanisms, we present in the Appendix an extremely short overview of the relevant material.

A critique of the optimality paradigm

The optimality paradigm may have particular relevance for isolated settings in which the network provider's goal is to maximize welfare, such as in a non-profit research network or an internal corporate network. In this paper, however, we are addressing the role of pricing in a commercially competitive environment. The current Internet service provision market has multiple independent service providers (ISPs), and competition appears to be increasing rapidly. We claim that the optimality paradigm is not an adequate foundation for pricing in such a competitive setting.

continued from page 183

[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. This research was supported in part by the Advanced Research Projects Agency, monitored by Fort Huachuca under contracts DABT63–94-C-0073 (SS), DABT63-94-C-0072 (DC), and DABT63-91-C-0001 (DE, SH). The views expressed here do not reflect the position or policy of the US government.

[2]See, for instance, 'Traffic jams already on the information highway' *New York Times* 3 November 1993 A1; 'Planet Internet' *Business Week* April 3 1995

[3]The authors include themselves in this critique, having adhered to this optimality paradigm in previous publications, such as Cocchi, R, Estrin, D, Shenker, S and Zhang, L 'A study of priority pricing in multiple service class networks' *Proceedings of SigComm '91* September 1991 123–130; and Cocchi, R, Estrin, D, Shenker, S and Zhang, L 'Pricing in computer networks: motivation, formulation, and example' *ACM/IEEE Transactions on Networking* 1993 **1** 614–627

The optimality paradigm places a special focus on marginal congestion costs. Our critique is posed in the form of three questions:

- Are marginal congestion costs relevant?
- Are marginal congestion costs accessible?
- Is optimality the only goal?

Are marginal congestion costs relevant?

It is a standard result that the overall welfare (the sum of provider profit and consumer surplus) is only maximized when prices are set equal to marginal cost, where these marginal costs take into account all externalities. In computer networks, these externalities include both congestion effects, where one user's use imposes a performance penalty on other users, and also connectivity effects, where a user benefits from other users being connected to the network.

Competition between network service providers will typically drive prices to these marginal costs. If the marginal cost prices are sufficient to recover the *facility* costs of building and operating the network infrastructure, then these marginal cost prices are a stable competitive equilibrium,[4] and so computing marginal congestion costs would be central to network pricing schemes. However, it is doubtful that such marginal cost prices will recover the full facility costs of computer networks. Within the context of the congestible resource model studied by MacKie-Mason and Varian,[5] marginal costs only cover the cost of the facilities priced at marginal expansion cost (ie the total congestion costs are equal to the product of the total capacity times the marginal cost of capacity). If the facility costs are a sublinear function of capacity [ie $f(x) > f'(x)x$ for $x > 0$], then facility costs will not be fully recovered by marginal cost pricing.[6] While the cost structure of networks is in flux as technologies rapidly evolve, it seems clear that a large portion of the facility costs arises from the fixed (ie not related to capacity) costs of deploying the physical infrastructure. Consequently, we assume in this paper that while marginal congestion costs may be non-trivial, they will be much less than the total facility cost of providing network service.

In such cases there is no stable competitive equilibrium;[7] any stable situation must have some prices that exceed the associated marginal costs. What guides the setting of prices in such a situation? While there are few general results applicable here, one could argue that the resulting prices of each firm will satisfy the Ramsey condition of maximizing the consumer surplus while still fully recovering costs (because otherwise competing firms would enter and lure customers away by offering more surplus). Thus, in raising prices to increase additional revenue, network service providers will do so in a manner that retains, to the greatest extent possible, the maximality of welfare (since maximizing welfare at a fixed level of profit is equivalent to maximizing consumer surplus at a fixed level of profit).

It is useful, in the following discussion, to artificially break the pricing structure into two distinct pieces.[8] One component of network charges is the attachment fee; this is the fee charged for gaining access to the network and is independent of any actual or potential usage. The other component is what we will call a usage-constraining fee. There are marginal costs associated with both attachment and usage, and welfare is optimized when these prices are set equal to their respective marginal costs. If all users derived significant benefit from their network connection, the Ramsey pricing scheme would be to raise attachment fees but

[4]We use the term *stable* only to mean that the revenues cover costs; we do not use the term to refer to any other dynamical properties of the equilibrium.

[5]MacKie-Mason, J and Varian, H 'Pricing the Internet' in Kahin, B and Keller, J (eds) *Public Access to the Internet* Prentice–Hall, Englewood Cliffs (1995) 269–314

[6]The seminal references on this phenomenon are: Strotz, R H 'Urban transportation parables' in Margolis, J (ed) *The Public Economy of Urban Communities* Resources for the Future, Washington, DC (1965) 127–169; and Mohring, H and Hartwitz, M *Highway Benefits: An Analytical Approach* Northwestern University Press, Evanston (1962)

[7]For a more thorough discussion of this point, see Srinagesh, P 'Internet cost structure and interconnection agreements' technical report. Bellcore (1995); Gon, J and Srinagesh, P 'The economics of layered networks' technical report, Bell Communications Research, Inc (1995)

[8]We ignore non-linear pricing policies here in order to simplify the discussion.

keep usage fees at the marginal cost levels, thereby retaining the optimal usage behavior and merely recouping additional revenue from attachment. This is the argument most commonly used to motivate the continued use of marginal congestion pricing in cases where marginal prices by themselves do not fully cover costs.[9]

Unfortunately, the assumption of uniformly large benefit from network access does not appear to apply to current computer networks. The low rates of penetration of Internet connectivity, and the high rate of churn in subscriptions to online services such as AOL and CompuServe, suggest that in addition to the many users that derive great value from their network connection, there are probably also many other users whose valuation of network connectivity is marginal and who would disconnect if attachment fees were raised.[10] Thus, we expect that both usage and attachment prices will affect welfare, and there will be a unique price point that produces a positive optimal welfare. Assuming smoothness throughout, deviating from the optimal pricing point produces welfare changes that are second-order in the price deviations. The matrix of second derivatives will depend in detail on the individual utility functions, and there is little reason to expect that, in general, consumer surplus is maximized when only attachment fees are raised. See the Appendix for some examples.

In addition, the Ramsey pricing scheme could be different for different subpopulations of users. For instance, low-volume users who derive very little benefit from being connected to the network would more likely absorb an increase in usage charges without detaching from the network. This is consistent with what we observe; some commercial Internet providers charge based on volume to attract low-volume, marginal benefit users who might not otherwise purchase access. In contrast, most large institutions, which typically derive great value from their network connection, pay substantially for the attachment. The traditional and cellular telephony markets also display extensive second-degree price discrimination (ie nonlinear pricing schemes where the per-unit price depends on the quantity purchased); there are many different pricing plans, some with lower attachment charges and higher usage charges, and others with the reverse. We expect a similar use of second-degree price discrimination to increase revenue in computer networks.

There are other considerations that suggest that usage charges must remain at significant levels, even if congestion is extremely low (and so marginal congestion costs are extremely low). If we assume that an entering network service provider can steal away a subpopulation of users from their current service provider if the entrant can supply this subpopulation with sufficient bandwidth to satisfy their needs at a cost less than the total fee being charged by the current provider,[11] then we must impose a 'core' condition on the pricing structure, mandating that no subset of users can be charged more than the cost of providing that subset service. If one believes that bandwidth is responsible for any significant portion of the cost of networks, then usage charges must be used to satisfy this core condition. Usage charges are needed to price discriminate between low-volume and high-volume users; otherwise a competing network provider would steal all the low-volume users away by offering a network provisioned at much lower levels with much lower prices. Thus, this core criterion requires that users who regularly consume (or who plan to consume) significantly less bandwidth be charged less, with

[9]We should make clear that we are assuming that network providers are not also controlling, or directly profiting from, the content delivered over their networks. However, in *bundled* networks such as cable TV, where the application and the network transport are sold as a single entity, there are many more opportunities to recover costs. Profits on content and revenue from advertising are important aspects of pricing in bundled networks. We do not consider such bundled networks in this paper but restrict ourselves to the analysis of pricing pure Internet access without bundled services. The nature of the ISP market is still very much in flux, and there may be other sources of revenue in the future, such as renting space on provider-supplied Web-servers, that may complicate the rather simplified case we are analyzing.

[10]It is possible that, in the grand and glorious future, the GII will have a single Internet-like network infrastructure and all households will have a single network connection that carries their telephony, television and data traffic. At that point, it may well be true that essentially all users have high valuation of their network connection and raising attachment fees is the appropriate way to raise revenue. However, we are a long way from this Utopian vision, and we should design our current network pricing policies to fit the present situation.

[11]This assumes seamless interconnection, so switching providers does not affect connectivity. Otherwise, the decision to switch providers involves many other factors besides cost.

the difference reflecting the percentage of cost due to bandwidth. Of course, if the bandwidth is relatively cheap (ie is a very minor portion of the network cost), then this 'core' argument has little bite.

Are marginal congestion costs accessible?

When prices are required to fully recover costs, we think there is little reason to expect that usage prices will equal the marginal congestion costs. We now put that conclusion aside and ask: if we nevertheless attempted to set prices to these marginal congestion costs, could we actually do so? It turns out that computing these congestion costs is quite difficult.

The relationship between what happens to a packet traversing a network and the resulting change in a user's utility is extremely complicated. When we look at the fate of a single packet, congestion can cause it to be delayed or dropped. Some applications are very sensitive to this extra delay (or being dropped), and others are not. Pricing schemes seeking to achieve optimal efficiency must take these different delay and drop sensitivities into account. While in simple theoretical models it is convenient to use the abstraction that a user's utility is a function of, say, average bandwidth and delay,[12] the real world is significantly more complicated.[13] Unfortunately, we have little beyond these simple theoretical models to guide us.

Moreover, most applications involve a sequence of packets, and the effect on utility due to the dropping or delay on one individual packet depends on the treatment given the rest of the packets. For instance, the performance of a file transfer depends on the time the last packet is delivered; for large files this transfer time depends almost exclusively on the throughput rate and not on the individual packet delays.[14] It is extremely difficult, if not impossible, for the network to infer the effect on the transfer time arising from delaying any of the individual packets, especially since the transfer time is also a function of the user's congestion control algorithm. To make matters even worse, often an entire suite of applications is used simultaneously, and then the user's utility depends on the relationship between the delays of the various traffic streams (eg a teleconference may involve an audio tool, a video tool and a shared drawing tool).

Our understanding of this relationship between handling of individual packets and the overall utility is rather primitive, and the relationship changes rapidly with technology (eg advances in congestion control could greatly decrease the sensitivity to randomly dropped packets). An important aspect of the problem is that the Internet architecture is based on the network layer not knowing the properties of the applications implemented above it. If we believe that network service providers will sell raw IP connectivity (ie they just provide access at the IP level and do not interpose any application-level gateways), then they have to price based solely on the information available at the IP level, and this greatly restricts the extent to which they can adjust prices to fit the particular applications being used.[15]

There have been many pricing proposals in the recent literature, and we do not attempt to review them all here.[16] The most ambitious pricing proposal for best-effort traffic is the 'smart-market' proposal of MacKie-Mason and Varian.[17] In this scheme, each packet carries a 'bid' in the packet header; packets are given service at each router if their bids exceed some threshold, and each served packet is charged this threshold

[12]As in, for example, Shenker, S 'Making greed work in networks: a game-theoretic analysis of switch service disciplines' *Proceedings of SigComm '94* August 1994 47–57

[13]See the discussion about the properties of best-effort traffic in Clark, D 'A model for cost allocation and pricing in the Internet' technical report, MIT (August 1995)

[14]See Clark *op cit* Ref 13 for a more thorough discussion of this point.

[15]The implications of this layering for content provision is discussed in Mackie-Mason, J, Shenker, S and Varian, H 'Service architecture and content provision' technical report, University of Michigan and Xerox PARC (1995)

[16]A few representative samples of the pricing literature are: Edell, R, McKeown, N and Varaiya, P 'Billing users and pricing for TCP' *IEEE Journal on Selected Areas in Communications* 1988 **13** 1162–1175; McLean, R and Sharkey, W 'An approach to the pricing of broadband telecommunications services' technical report, Bellcore (March 1993); Stahl, D and Whinston, A 'An economic approach to client-server computing with priority classes' technical report, University of Texas at Austin (1992)

[17]MacKie-Mason, J and Varian, H 'Pricing the Internet' in Kahin, B and Keller, J (eds) *Public Access to the Internet* Prentice-Hall, Englewood Cliff (1995); MacKie-Mason, J, Murphy, J and Murphy, L 'ATM efficiency under various pricing schemes' technical report, University of Michigan, Dublin City University and University of Auburn (March 1995); MacKie-Mason, J, Murphy, J and Murphy, L 'The role of responsive pricing in the Internet' technical report, University of Michigan, Dublin City University and University of Auburn (June 1995)

price regardless of the packet's bid. This threshold is chosen to be a market clearing price, ensuring the network is fully utilized. The threshold price can be thought of as the highest rejected bid; having the packets pay this price is akin to having them pay the congestion cost of denying service to the rejected packet. The key to this proposal is incentive compatability; users will put their true valuation in the packet since, as in standard second-price auctions, it only effects whether they get service but not how much they pay. By putting their true valuation of service in the packet header, users will get service if and only if it costs them less than their valuation of the service.[18]

This proposal has stimulated much discussion and has significantly increased the Internet community's understanding of economic mechanisms in networks. However, there are several problems with this proposal that prevent it from achieving true optimality. First, the most fundamental problem is that submitting a losing bid will typically lead to some unknown amount of delay (since the packet will be retransmitted at a later time), rather than truly not ever receiving service, so the 'bid' must reflect how much utility loss this delay would produce rather than the valuation of service itself; thus, accurate bids cannot be submitted without precisely knowing the delay associated with each bid level, and neither the network nor the user knows this delay. Second, there are complications when the packet traverses several hops on its way to its destination. The valuation is an end-to-end quantity (the user only cares about the packet reaching its final destination and does not care about any partial progress), yet the valuation is used on a hop-by-hop manner to determine access at each hop; one would have to extend the bidding mechanism to evaluate the entire path at once, and this entails a distributed multiple good auction of daunting complexity.[19] Third, the bid is on a per-packet basis, yet many applications involve sequences of packets. It is impossible to independently set the valuation of a single packet in a file transfer, when the true valuation is for the set of packets.

Wang *et al*[20] have proposed a pricing scheme for flows making network reservations (ie asking for a quality of service that entails admission control and some assured service level) where prices optimize a given objective function. Gupta *et al*[21] adopt a similar approach for a best-effort network with priorities. As in any conventional economic setting, the optimality of the pricing scheme depends on knowing the demand function. In settings where the supply and delivery are not time critical, such demand functions can be estimated over long periods of time. However, in computer networks, a user's utility depends on the delay in meeting his/her service request, and so one cannot merely consider the long-term average demand but must also respond to instantaneous fluctuations in demand when setting prices. In addition, the problem of denial of service leading to some delay, rather than an eternal denial of service, makes the valuations of the flows not directly related to congestion costs. Consequently, determining optimality in the presence of fluctuating demand is extremely difficult.

We contend that the failure of these mechanisms to achieve true optimality is not a failure of imagination, but rather evidence that the task is beyond the scope of any practical algorithm. The key to efficiency—knowing the service degradation that will result from a particular network action (ie how much delay and/or loss), and knowing the user's utility loss as a result of this service degradation—is fundamentally unknowable.

[18]As an aside, note that the pricing scheme is embedded within the architecture in this proposal. The bids are translated into the prices charged.

[19]If one believes that the major source of congestion is at the edge of the network, then one could apply the smart market only at the edge points. This removes the end-to-end versus per-hop problem and could be used in our edge pricing scheme as the method of charging. See the discussion in the section titled 'Forms of pricing'.

[20]Wang, Q, Sirbu, M and Peha, J 'An optimal pricing model for cell-switching integrated services networks' technical report, Carnegie Mellon University (May 1995)

[21]Gupta, A, Stahl, D and Whinston, A 'Managing the Internet as an economic system' technical report, University of Texas at Austin (July 1994)

This is not to imply that usage pricing schemes are of little utility. When compared to a situation with no usage-constraining charges, usage charges greatly increase the efficiency of the network. Simulations and calculations[22] have clearly demonstrated the significant advantages usage pricing has over free entry. Our point is merely that such pricing schemes do not achieve true optimality, and that the significant efficiency gains demonstrated could probably also be achieved with explicitly suboptimal schemes.

Is optimality the only goal?

We argued in the previous section that marginal congestion costs are inherently inaccessible. This critique applies equally to attempts to compute the optimal Ramsey prices. However, since price deviations away from these optimal points typically produce only second-order deviations in the total welfare, perhaps such deviations are not of much concern. Moreover, in the pursuit of optimality in simplified models, some more basic structural issues have been somewhat neglected. In this section we identify some of these structural issues and urge that they be given significant attention in the design of pricing policies.

Pricing policies should be compatible with the structure of modern networking applications. One of the recent developments in the Internet is the increasingly widespread use of multicast, in which a packet is delivered to a set of receivers, rather than just a single receiver. By sending packets down a distribution tree, and replicating packets only at the tree's branch points, multicast greatly reduces the load on the network. Therefore, it is crucial that pricing give the proper incentives to use multicast where appropriate.

Another important aspect of network applications is that the benefit of network usage sometimes lies with the sender of the traffic and sometimes with the receiver(s). Pricing mechanisms should be flexible enough to allow the charges to be assessed to either, or some combination of both, end-points. This is a very important goal in computer networks; the ability to charge receivers would facilitate the free and unfettered dissemination of information in the Internet, since the providers of such information would not have to pay the cost of transport. Note that this goal is not achieved by the flat pricing approach; currently the source's access charge is paid exclusively by the host institution. This has not yet caused a problem on the Internet, since the elastic and adaptable data applications can easily adjust to overloaded conditions. However, when real-time applications, and other applications that adapt less well to congestion, are in widespread use the pinch at the source's access point will be felt more acutely.[23]

Pricing policies should also be compatible with the structure of the network service market. There are numerous independent service providers, and many of these are small providers who merely resell connections into bigger provider networks. The interconnection arrangements between providers are somewhat *ad hoc* and changing rapidly.[24] Interconnection among these networks is crucial for maximizing social welfare. Pricing schemes should not hinder interconnection by requiring detailed agreement on pricing policies and complicated per-flow transfers (ie a separate transfer for each flow) of money when carrying traffic from another interconnected network. In addition, these independent service providers should be able to make local decisions about the appropriate pricing policies. This implies that the pricing

[22]Gupta, A, Stahl, D and Whinston, A 'The Internet: a future tragedy of the commons' technical report, University of Texas at Austin (1995); MacKie-Mason, J and Varian, H 'Pricing congestible network resources' *IEEE Journal on Selected Areas in Communication* 1995 **13** 1141–1149; MacKie-Mason, J, Murphy, J and Murphy, L 'The role of responsive pricing in the Internet' technical report, University of Michigan, Dublin City University and University of Auburn (June 1995)

[23]The ability to assign charges to the receiving end could, in some cases, be handled by a higher level protocol that redistributes the basic charges determined by the network. However, there are several disadvantages to requiring such a high level protocol: it requires the ability to transfer funds at a higher level, it cannot deal with capacity-based charging, and in the multicast case the required information (such as the membership of the group and the network topology) may not be available at the higher layer. Thus, we think it preferable to build the flexibility of assignment into the basic charging mechanism itself.

[24]Srinagesh *op cit* Ref 7; Gon and Srinagesh *op cit* Ref 7

policy should not be embedded into the network architecture. Instead, the network architecture should provide a flexible accounting infrastructure that can support a wide variety of locally implemented pricing schemes. For instance, there are some contexts (such as managing an internal corporate or university network) where the goal of pricing is merely to encourage efficient use of the network resources. Often in these contexts there are incentives that can be used (eg quotas) instead of money. While in this paper we have focused on monetary incentives, the underlying accounting structure and pricing architecture should allow the use of these other incentive forms if they are locally applicable.

Note that achieving optimality necessarily involves uniform implementation of a single pricing scheme across the network; optimality involves setting prices at exactly the marginal congestion costs, and so the accounting scheme becomes a distributed computation of those congestion costs. Thus, the optimality paradigm is fundamentally inconsistent with the need for locality in pricing. Given that no pricing scheme claims to be truly optimal, the need for local control should take precedence over the desire for absolute optimality.

While true optimality is not an appropriate goal, pricing should still be used to achieve reasonable levels of efficiency. It is important that the underlying accounting infrastructure allow prices to be based on some approximation of congestion costs. There is an important distinction lurking here. It is important to allow prices to be *based* on some approximation of congestion costs, but it is important not to force them to be *equal* to these congestion costs. As we argued, the need for full cost recovery militates against such an assumption of equality. Meeting any reasonable efficiency goal, however, would likely require that prices depend on such congestion costs.

Rather than start with mechanisms designed to precisely calculate marginal congestion costs, we might first ask: what are the absolutely minimal requirements for providing some estimate of congestion costs? One minimal requirement is that pricing should encourage the appropriate use of quality of service (QoS) signals (by this we mean the signals sent by applications to the network requesting a particular quality of service; see Appendix). This is crucial for making the new QoS-rich network designs effective and would enable them to achieve significant increases in network efficiency. An additional requirement is that pricing should discourage network usage during times of congestion, but not discourage it during relatively uncongested times. Our basic point is that perhaps these minimal requirements are sufficient to achieve reasonable approximations, and that attempts to more accurately calculate Ramsey prices are of little (indeed second-order) value and distract us from the more important but often overlooked structural concerns.

A new pricing paradigm: edge pricing

After having critiqued the reigning optimality paradigm, we now present a very different pricing paradigm: edge pricing. We motivate the edge pricing paradigm by describing a series of approximations to true congestion costs.

Approximating congestion costs

Computing the true congestion costs requires that you can compute

other users' loss in utility due to one user's use. This requires knowledge not only of the utility of users, which in the Internet architecture is fundamentally unknowable, but also the knowledge of the current congestion conditions along the entire path. Such detailed knowledge entails a sophisticated accounting scheme that transcends administrative boundaries by following the entire path. Having already concluded that our estimates of utility loss are extremely rough estimates, can we also replace the knowledge of current congestion conditions along the entire path with a reasonable, but more easily accessible, estimate? Consider the following two approximations.

The first approximation is to replace the current congestion conditions by the *expected* congestion conditions. This is essentially QoS-sensitive time-of-day pricing. The time-of-day dependence builds in expectations about the current congestion conditions. The QoS dependence reflects the fact that the effect one flow's packets have on another flow's packets depends on the respective service classes of the flows; packets in higher quality service classes impose more delay on other packets than do packets in lower quality classes. This approximation of QoS-sensitive time-of-day pricing has the problem that it does not reflect instantaneous fluctuations in traffic levels; packets sent during a lull in the network would still be charged full price even though the actual congestion costs were quite small. Such insensitivity to instantaneous conditions would seem to remove any incentive for users to redistribute their load dynamically; just as in the telephone network, time-of-day pricing encourages users to time-shift their calls to later (or earlier) hours when rates are lower, but does not encourage them to adjust to the instantaneous conditions. (Of course, in the telephone network there is no way for users to detect the current load.)

We claim that the inability to charge less during periods of low congestion is not a serious problem because, in many cases, one can substitute the congestion-sensitivity of service for the congestion-sensitivity of prices. During a lull in the network, lower quality classes give as good service as high quality classes during congested periods. Users who monitor the service they are getting from the network and adjust their service request accordingly can take advantage of this variability. The way user costs are lowered during times of reduced network load is not that the network lowers the price of service classes but that users request lower service classes and are charged the lower price of that class.

We refer to 'users' as the entities adapting to current conditions, to distinguish this from the network adapting, but we should note that in reality adaptation does not require significant effort from the human user.[25] Instead, adaptation routines will be highly automated and embedded within applications or the end system's operating system. Many current network applications are already designed to adapt to network conditions, and so relying on users to adapt to current conditions, rather than the network, is quite consistent with current practice. In fact, this reflects a basic Internet design philosophy; to the extent possible (and routing is the one place where it is frequently less possible), the intelligence and responsibility to adapt to current conditions should be placed on the outside of the network; the fundamental infrastructure inside the network should remain fairly simple, intentionally ignorant of the applications it is supporting, and should not try to adapt on behalf of these applications. Applied to this case, this

[25]See MacKie-Mason, Murphy, and Murphy *op cit* Ref 22 for a similar discussion of the role of adaptation.

philosophy argues for relatively static pricing policies with end-users varying their service requests in response to current congestion conditions. This removes from the network the responsibility of accurately assessing current conditions and their likely impact on users' utilities and puts the onus on individual applications/users to make that assessment for themselves; given that applications have very different sensitivities to service quality, it seems preferable to place the bulk of the variability where it can be done in the most informed way.

If expected congestion were the only approximation, then we would essentially have a pricing scheme where prices were computed per-link based on the time-of-day and quality of service requested. The second approximation is to replace the cost of the actual path with the cost of the *expected* path, where the charge depends only on the source and destination(s) of the flow and not on the particular route taken by the flow. From a user's perspective, they have requested service from one point to another (at least in the unicast case); the actual path the data takes is typically determined by the networking routing algorithms (except in the case of source routing). Having the price of the service depend on the network's decision about routing seems an unnecessary source of price variation that makes it harder for the user to make informed plans about network use. Moreover, when alternate paths are taken by the network in response to congestion, the extra cost due to the congestion should not necessarily fall only on those flows that have been redirected. Certainly in the telephone network, the price of a telephone call does not depend on the network's choice of route.

Edge pricing

When we combine these two approximations, the price is based on the expected congestion along the expected path appropriate for the packet's source and destination. Therefore, the resulting prices can be determined and charges assessed locally at the access point (ie the edge of the provider's network where the user's packet enters), rather than computed in a distributed fashion along the entire path. We will call this local scheme *edge* pricing. A similar approach to pricing in computer networks has been suggested by Jacobson.[26] The prices charged at the edge, or access, point may depend on information obtained from other parts of the network, but the entire computation of charges is performed at the access point. In the 'Architectural issues' section we discuss the multicast case where the relevant information is difficult to obtain.

As discussed by Clark,[27] edge pricing has the attractive property that all pricing is done locally. Interconnection here involves the network providers purchasing service from each other in the same manner that regular users purchase service. When a user connected to provider A's network sends a packet, it is applied to that user's bill according to whatever pricing policy provider A has.[28] If the destination of the packet is on provider B's network, then when the packet enters provider B's network the packet is charged against provider A's bill with provider B. There are no per-flow settlement payments, in the sense that the various providers do not redistribute the charge levied to the end-user among themselves. Instead, each provider takes full responsibility for every packet they forward; a sequence of bilateral agreements between the adjacent service providers along the path performs the necessary function of cost-shifting. These bilateral agreements apply only to the aggregate usage by these providers and thus greatly simplify the transfer

[26]Jacobson, V, private communication (1995)

[27]Clark *op cit* Ref 13

[28]We use the term 'bill' here only to connote that the packet is applied to the contract the user has with provider A; as we mention below, the contracted pricing policy may very well be a flat price with a limit on peak rate, in which case there is no additional charge per packet.

of payments between providers.

The beauty of this is that billing structures are completely local. The exact nature of the pricing scheme is simply a matter between the user and the service provider. Because the decisions are local, service providers can invent ever more attractive (and complicated) pricing schemes and can respond to user requirements in a completely flexible fashion. No uniform pricing standards need to be developed, since interconnection involves only bilateral agreements that allow each provider to use their own pricing policy. Locality allows providers to experiment with new pricing policies and gradually evolve them over time; in fact, pricing policies will likely be one of the important competitive advantages available to providers when competing with each other. For instance, locality allows providers to offer specialized pricing deals such as bulk discounts. It is hard to imagine implementing a meaningful bulk discount when charging is done in a non-local per-link basis; a user's usage of any particular link, or of any particular service provider outside of the local one, is probably quite limited, and so such discounts are much less meaningful.

Forms of pricing

Edge pricing describes the place at which charges are assessed but is completely neutral about the nature of these charges. In most of the literature, there is a sharp distinction between usage and attachment charges; this differentiates the fixed (or flat) portion of the price and the variable usage-dependent portion of the price. Thus, the cost of upgrading the speed of a user's access line is considered an attachment charge. We think this division is somewhat misleading, since there is a natural continuum between the two.[29] We instead choose to refer to them all as usage-constraining prices.[30] Per-packet charges are clearly designed to constrain usage, but so are limits on a user's peak sending rate.

The continuum of usage-constraining charges can perhaps best be explored by defining its two end-points. At one end of the continuum, prices can be based on actual usage, in the form of per-packet and/or per-reservation charges; this is the traditional form of usage-based pricing. At the other end of the spectrum, users could purchase a capacity from the network and then be allowed to use, without any additional charge, up to that capacity. One form of capacity could be defined in terms of just a peak rate, as in the current form of flat pricing. More generally, however, this capacity is defined in terms of a *filter* that is applied to the traffic. A usage filter characterizes flows as either conforming or not conforming to the agreed upon capacity. Such filters can measure the usage over differing time horizons, such as controlling the long-term average rate, the short-term peak rate and intermediate burst durations. This capacity framework is merely a generalized version of the current flat-rate pricing schemes; the extra flexibility allows pricing schemes to be more closely attuned to user requirements. (See the Appendix for a more complete explanation of such filters.) While not essential to our discussion here, we should note that there can be several possible actions that the service provider could take when a user exceeds his/her capacity; for instance, all such packets could be mapped into the lowest service class, or dropped, or queued until the flow is in compliance with the filter, or merely assessed an additional per-packet fee.

[29]The distinction between fixed and variable prices may be extremely important to individual users; users on fixed budgets may need fixed prices, whereas users with extremely variable demand may need the ability to only pay for usage. Our point is that this distinction, while important to individual users, is not fundamentally important from an architectural or economic perspective. Both forms of pricing can be assessed locally, and both constrain usage.

[30]In our taxonomy, attachment prices would refer only to the price of attaching to the network and not refer at all to the speed of the access line. All other charges would be considered usage-constraining. Of course, in non-linear pricing schemes (or when there is a spectrum of pricing menus offered, as in the current cellular telephony market) the distinction between the two is completely blurred.

The units of usage that are applied against the capacity constraints, just like per-packet charges, can depend on many things such as time-of-day, destinations and QoS. High quality service classes might consume twice as many units as lower quality service classes, with similar increments for packets traveling further or over particularly congested links. Of course, to realize the goal of allowing users to send an unlimited amount of traffic when the network is empty, there should be a category of absolutely lowest quality of service that is essentially free. In fact, one could even use a 'smart-market' auction approach to pricing at the access point.[31]

These capacity constraints allow network providers to make informed provisioning decisions. Of course, provisioning decisions will also be heavily based on measurements of actual aggregate usage, but the capacity filter parameters give some additional input for estimates. If there is an infinite amount of multiplexing (ie each user constitutes an infinitesimal share of the aggregate usage) and users are uncorrelated, then provisioning need only be based on the long-term average rates. The other capacity filter parameters are needed to make estimates of the magnitude of usage fluctuations away from this average value.

Because the overlimit behavior (when usage exceeds the capacity) can merely be an additional per-packet charge, there can be a continuum of pricing policies that stretch between purely usage-based charging and purely capacity-based charging. Within the spectrum of edge charging, the difference between capacity-based prices and usage-based prices is not a fundamental architectural issue. We expect that the market will invent, over time, increasingly attractive and flexible hybrids of these approaches. Telephony may provide an instructive example. Telephone companies offer a menu of local calling plans, some usage-based (eg metered service), some capacity-based (eg unlimited service) and some a combination of both (eg a certain number of free minutes per month, plus a metered rate for calls in excess of this number). It is likely that the same will happen in computer networks, with some users choosing usage-based and others choosing capacity-based charges, and many choosing something in between. Thus, the heated debate between advocates of usage-based and capacity-based pricing schemes will become completely irrelevant as users vote with their feet. Because in the edge pricing paradigm the decision between usage-based and capacity-based, or something in between, is completely local, and we expect that network provision will be competitive, the offererd plans will likely reflect the true needs of consumers (and thus the architecture need not preclude one choice or the other to prevent providers from exploiting users).

The rest of this paper is devoted to exploring the infrastructure needed to support this edge pricing approach.

Architectural issues

Edge pricing localizes the whole charging process; everything occurs at the access point. Yet, there are two inherently non-local aspects of pricing: (1) charging appropriately for multicast, and (2) the ability to charge receivers for the service.[32] These non-local aspects pose some fundamental architectural challenges to the edge pricing approach. We see these issues as forming the basis of a fertile research agenda in pricing in computer networks. In this section we discuss how the

[31]There may be some disadvantages with using the smart-market as the local pricing scheme (eg it embeds the pricing policy in the architecture), but our point here is that it is not architecturally precluded by the edge pricing paradigm, and so firms are free to experiment with it.

[32]Charging receivers for service is a non-local problem because, in approaches with explicit willingness-to-pay signaling, when both the source and receiver are serviced by the same provider the source's access point must be informed that the receiver is willing to assume responsibility for the transmission. Similarly, when the path from source to receiver traverses several different provider networks, the notification of receiver-paying must be communicated to both the exit access point and the entrance access point in each network. Other approaches can avoid this explicit nonlocal signaling by adopting some uniform standards, such as a certain portion of the multicast address space being set aside for receiver-pay groups, but these standards themselves are nonlocal in that they represent agreements between providers about a billing policy.

infrastructure might be designed to handled these non-local aspects. We describe the problems of multicast and charging receivers separately, and then we review some remaining open problems.

It is important to note that this design discussion is extremely preliminary and is intended to be illustrative rather than definitive. That is, our purpose is to illustrate some of the issues involved by engaging in a design discussion, but we freely admit that the design directions advocated here may not, in the end, be the appropriate choices.[33]

Multicast

When unicast packets enter a provider's access point, the destination field is enough to determine the typical path of the packet. Unicast routes fluctuate occasionally, but the normal case is that unicast routes change on rather slow time scales. Thus, fairly static tables at the entry points can provide adequate information for pricing decisions, and it would be relatively trivial to design the distributed algorithms needed to construct and maintain these tables.[34] If addresses encode geographic information (as in Deering's recent proposal[35]) or provider information (as in the current IPv6 proposal[36]), then these tables are especially simple.[37] Moreover, if the provider networks are small enough, one set fee for all intraprovider packets and another fixed fee for all interprovider packets might be sufficient.

Multicast packets pose more of a challenge. A multicast address is merely a logical name, and by itself conveys no geographic or provider information. While multicast routing identifies the next hop along the path for packets arriving at an interface, multicast routing does not identify the rest of the tree. Thus, estimating costs in the multicast case requires an additional piece of accounting infrastructure. Moreover, the set of receivers—the members of the multicast group—can change quite rapidly, and so the mechanisms for providing the appropriate accounting information must be designed with care.

One can imagine several different approaches. The simplest would be to merely collect the location (ie subnet numbers) of all receivers (with receivers outside of the provider's network being recorded as residing at the appropriate exit point of the network). From these locations one could compute the approximate costs of the appropriate tree.

Another approach would be to compute these costs on-the-fly by introducing a new form of control message—an accounting message—that would be initiated when the receiver sends its multicast join message (multicast *join* messages are the control messages sent by a receiver to join the multicast group; see Appendix). These accounting messages would be forwarded along the reverse trees towards each source, recording the 'cost' of each link it traversed and summing costs when branches merged. When these accounting messages reached a source's access point, the cumulative cost of reaching all receivers from that source would be available. Each provider would only need to record the cost information local to their network; that is, the costs would start accumulating when the accounting message entered the provider's network and would stop when the accounting message exited the network. No cost information crosses the provider boundaries; instead, this cost information is only used locally to compute the charges to apply on the edge of the network. This on-the-fly approach makes the charge for multicast dependent on the true path rather than the typical path, which may cause unnecessary variability.[38] We should note that

[33]In particular, there is a spectrum of design choices providing different levels of functionality and requiring more or less additional mechanism; in the following discussion we are not attempting to make a detailed evaluation of the functionality versus mechanism tradeoff, but are merely illustrating some possible ways of achieving the aforementioned goals. There are more minimalist approaches to these problems that require less additional mechanism, and they should be considered when making design decisions for the Internet, but for this pedagogical discussion we have presented more straightforward, if more mechanistic, approaches.

[34]These tables would contain information describing how many usage units (for a capacity filter) each packet represents, or a monetary per-packet charge, or whatever other information is needed for the provider to assess the appropriate charges.

[35]Deering, S, private communication (1994)

[36]Deering, S and Hinden, R 'Internet protocol, version 6 (ipv6) specification' Internet draft (June 1995)

[37]For more information on addressing options, see Francis, P 'Comparison of geographical and provider-rooted Internet addressing' *Proceedings of INET '94* June 1994

[38]One could apply such a scheme to a logically overlaid network so the prices would be less dependent on the details of the path. For instance, the network could be divided up into area codes, with logical link costs recorded whenever the accounting message left one area code and entered another.

there might be groups (eg cable TV channels) where the typical tree might be well enough known in advance so that such additional mechanisms are not needed.

Note that the additional piece of accounting infrastructure needed to compute these costs is local to the provider; that is, each provider can use its own algorithm. No standards need be established and no agreements with other providers need be made. Thus, this protocol can incrementally evolve over time as we understand better the cost structure and traffic patterns of future networks. Independent evolvability is one of the biggest advantages of the edge pricing paradigm; while the total amount of mechanism needed to perform the necessary accounting may not be less than in other paradigms, the degree of independence of these accounting mechanisms is substantially higher. The ability for providers to act independently to upgrade their accounting will lead to rapid development of the required implementations; proposals that require a single uniform and standardized accounting infrastructure are much less likely to ever be implemented.

The above discussion applies to the complete spectrum of usage-constraining pricing schemes, from usage-based to capacity-based charging. However, much of the above discussion implicitly applied to best-effort service. The basic principles remain the same when pricing for reserved or assured levels of service, but the mechanistic details are quite different because of the presence of a set-up protocol like RSVP.[39]

Charging receivers

The second non-local problem we consider is assigning charges to receivers. This involves addressing the following three issues.

(1) *How does a receiver indicate to the network provider that it is willing to take responsibility for the source's traffic?* Here there are several alternatives, and we merely mention a few to illustrate some of the possibilities. In the best-effort multicast case, the join message might be extended to include a willingness-to-pay field. In the case of reservations, either unicast or multicast, the RSVP reservation message could carry similar information. The only case that does not already have a pre-existing control message that could be used for this purpose is the unicast best-effort case. Here, we may require a new willingness-to-pay control message to be generated by the receiver, but there may also be other approaches. In addition, we may want to allow the source to indicate that it is not willing to pay, so that if the source's access point has not received a notification that the receiver(s) is (are) willing to pay, then the packets are immediately dropped; such an indication that the source is not willing to pay could be contained in the packet header. Another approach—one that requires no additional signaling—is to divide the multicast addresses into sender-pays and receiver-pays categories, so that the assignment option is indicated by the choice of multicast address. Here the very act of joining the group communicates a willingness to pay.

(2) *How does the network 'bill' the receiver?* One general approach here is to apply pricing when the packet's traverse the receiver's access point. Thus, packets are 'charged' according to the receiver's contract with its provider, not according to the sender's contract. If the capacity is exceeded, then the overlimit behavior (delaying, dropping, etc) is applied at this exit point. If the packet traverses several providers, then this reverse charging is applied whenever a boundary is crossed; the

[39]Zhang, L, Deering, S, Estrin, D, Shenker, S and Zappala, D 'RSVP: a resource reservation protocol' *IEEE Network Magazine* 1993 **7** 8–18

packet is charged to the provider whose network the packet is entering, not the provider whose network the packet is exiting.

(3) *How does the network split the responsibility for the bill among the members of a multicast group?* If there are multiple receivers, the network not only needs to transfer the charges to the receivers, but also must apportion the cost among them in a reasonable manner. One way to do this is to assign fractional responsibilities to each of the receivers. Then, when the packet arrives at each receiver's access point, the receiver is 'charged' only the fraction of the normal amount. The variety of policies for assigning these fractions, as well as mechanisms for computing them, have been addressed in previous work.[40] One could also use cruder approximations to compute these fractions, basing multicast prices on *ad hoc* discounts from the unicast cost.

Open issues

The preceding general discussion merely presents some possible approaches. There are many other possibilities, and the ones mentioned above should be considered sketchy illustrations of the issues involved, rather than serious and complete design proposals. This initial design discussion leaves several fundamental issues unresolved; we mention a few of these here.

Our discussion of the need for charging receivers has focused on a narrow binary choice; either sources pay or receivers pay. One may want to consider a much broader spectrum of policies in which the costs are shared in a more flexible manner. This might be a fractional splitting (eg the source pays 30% and the receiver pays 70%), or perhaps the source pays for a certain portion of the path (eg the source pays for the portion of the path within its local provider's network) and the receiver pays the rest. We have not addressed the requirements of such source/receiver cost sharing.

We also have not considered the case where some receivers are willing to pay and others are not. Aside from the mechanistic questions, there are important unresolved policy questions about how to handle such a situation. Related to this is the fact that receivers may want to limit their exposure. The willingness-to-pay field may, in addition to indicating that the receiver is willing to pay, also indicate a cap on how much (in some arbitrary units) cost the receiver is willing to absorb. For instance, when a receiver in California joins a group for a virtual rock concert sourced from London with an expected audience of millions, the receiver may be willing to pay his/her share of a few dollars (or equivalent capacity units) but would certainly not be willing to absorb the bill for the entire 50 mbps video feed from London. However, such limits open up thorny strategy issues, as receivers would be tempted to *free ride* on other receivers.

There may be other approaches to deal with the startup phase of multicast groups that will eventually become large. There may be some way that the organizer of the session, whether it be a rock concert or an IETF broadcast or a cable TV channel, could describe to the network beforehand an approximation of the likely distribution tree. This would enable the network to estimate the likely cost shares beforehand and thus greatly reduce the exposure of the first few group members. Other schemes to reduce such exposure have been discussed by Gawlick.[41]

The accounting mechanisms discussed by Herzog *et al*[42], which determine the appropriate multicast cost shares, are implemented on a

[40]Herzog, S, Shenker, S and Estrin, D 'Sharing the "cost" of multicast trees: an axiomatic analysis' *Proceedings of SigComm '95* August 1995 315–327

[41]Gawlick, R 'Admission control and routing: theory and practice' technical report, MIT (August 1995)

[42]Herzog *et al op cit* Ref 40

link-by-link basis. Such methods must be extended to a more abstract set of logical links so that the cost shares can reflect a coarser level of granularity. Also, some cost-sharing approaches depend on the number of receivers downstream of each link. Such numbers are relatively easy to obtain within a provider's network (eg by extending the multicast join mechanism). To do this accurately across providers would require each provider to reveal this number to other providers, and this raises an incentive question, since the cost share increases with the number of receivers (and so each provider would reveal only the existence of one receiver in their network).

Another issue arises in the case of receiver-pays with capacity-based charging, where the overlimit behavior is packet dropping. If the incoming traffic greatly exceeds the available capacity, then the network has transported packets across the network only to consistently drop them at the exiting access point. We may need some slow, out-of-band signaling in this case to *unjoin* the receiver from the group. Such signaling is not needed if the overlimit behavior is an additional per-packet charge.

Conclusion

Current discussions about pricing in computer networks are dominated by two main topics. The first topic is the debate between usage-based pricing and flat pricing that has embroiled the network design community and caught the attention of the popular press. Rather than being radically different, we think these two schemes reside along the single continuum of *usage-constraining* pricing policies. As in telephony, both pricing options, along with various intermediate hybrids, will likely be offered to users by their local provider. The detailed design of such schemes is perhaps best left to the marketing departments of the various network service providers. Thus, no particular pricing policy should be embedded into the network architecture. The challenge for the network design community is to provide a coherent network pricing architecture that allows individual providers to make their own choice about how to price service. This paper presents one such pricing architecture that achieves this goal: *edge* pricing.

The second topic, emphasized in the more academic literature, is the design of marginal cost pricing schemes that produce the optimally efficient use of network resources. We have critiqued this optimality paradigm on three grounds: (1) marginal cost prices may not produce sufficient revenue to fully recover costs and so are perhaps of limited relevance, (2) congestion costs are inherently inaccessible to the network and so cannot reliably form the basis for pricing, and (3) there are other, more structural, goals besides optimality, and some of these goals are incompatible with the globally uniformity required for optimal pricing schemes. For these reasons, we contend that the research agenda on pricing in computer network should shift away from the optimality paradigm and focus more on structural and architectural issues. Such issues including allowing local control for pricing policies, fostering interconnection, handling multicast appropriately, and allowing receivers to pay for transmission. To illustrate our point, we described how these goals might be accomplished in the context of the edge pricing paradigm.

Even though many of our detailed comments concern the particular

pricing architecture of edge pricing, our intent in writing this paper is not to advocate that this scheme be adopted by the Internet; the proposal is extremely preliminary and there may be other schemes with similar properties. Rather, our intent is to initiate a dialogue about such pricing schemes and hopefully stimulate the creation of other pricing paradigms that meet our design goals.

Appendix

Internet architecture and mechanisms

This section describes a few relevant features of the Internet architecture. It is a very selective and sketchy overview, intended merely to provide a minimal background for reading this paper. The current Internet architecture is designed for point-to-point (or unicast) *best-effort* communication. Every packet header contains a source address and a destination address. Upon receiving a packet, a switch (or, equivalently, a router) consults its routing table to find, based on the packet's destination address, the appropriate outgoing link for the packet. Sometimes the incoming rate of packets at a switch is greater than the outgoing rate, and so queues build up in the switch. These queues cause packet delays and, if the switch runs out of buffer space, packet discards. The network does not attempt to schedule use; sources can send packets at any time, and the network switches merely exert their *best effort* to handle the load.

This simple network architecture has been amazingly successful. However, there are efforts currently underway to extend this architecture in two ways. The first is to offer better support for multipoint-to-multipoint communications through the use of *multicast*.[43] In the current Internet architecture, when a source sends a packet to multiple receivers, the source must replicate the packet and send one to each receiver individually. This results in the several copies of the same packet traversing those links common to the delivery paths (ie those links that lie on more than one delivery path, where the delivery path is the route taken by the packet from source to receiver). In multicast, the source merely sends the packet once, and the packet is replicated by the network only when necessary (ie the packet is transmitted only once on each link, and then is replicated at the split points where the delivery paths diverge and one copy is sent along each outgoing branch). Sources sending to a multicast group use the multicast group address as the packet's destination address; however, multicast addresses are merely logical names and do not convey any information about the location of the receivers (unlike a unicast address). Computers on a network wishing to receive packets sent to a particular multicast address send a *join* message to the nearest router. The routing algorithm then distributes this information to create the appropriate distribution trees (ie trees from every source to every receiver) so that packets sent to the group reach each receiver. There is a variety of routing algorithms that can accomplish this task, and we do not review them here. Note that senders are not aware of who is receiving the packets, since the multicast paradigm is receiver-driven. Efforts to standardize and deploy multicast are well advanced; the vitality of the current MBone[44] attests to the benefits of this technology.

The second extension to the Internet architecture is much more preliminary and rather controversial. Efforts are underway to extend the Internet's current service offerings to include a wider variety of qualities of service (QoS). The current single class of best-effort service may not be sufficient to adequately support the requirements of some future video and voice applications (although this is a highly debatable point[45]). Moreover, offering all applications the same service is not an efficient use of bandwidth; providing a wider variety of qualities of service allows the network's scarce resources to be devoted to those applications that are most performance-sensitive. There are many ways in which these services could be extended, some as simple as merely providing several service priority levels and/or drop priority levels. See also the discussion by Clark[46] for other approaches to such extensions to best-effort service. Offering multiple qualities of service requires some form of incentives, such as pricing, to encourage the appropriate use of the service classes.[47]

More radical extensions to the service offerings are also contemplated. A working group of the Internet Engineering Task Force is preparing a proposal to offer several *real-time* services; a bounded-delay service, in which the network commits to deliver all packets within a certain delay, is an example of such a real-time service. These services are fundamentally different than best-effort in that the network is making an explicit and quantitative service commitment and therefore must reserve the appropriate resources. Such services require admission control procedures, whereby receivers request service (ie issue a reservation request), and the network then either commits to the requested level of service (if it can meet the requirements) or denies the reservation request (if the current load level is too high to meet the requirements of the incoming request). In the proposed resource reservation protocol, RSVP, receivers send their request for

service to the network, and this request follows the reverse delivery tree towards all relevant sources (a single source if the application is unicast, or to all senders in the group if the application is multicast).[48]

Ramsey prices in a simple model

In this section we explore the behavior of Ramsey prices in a simple network model. We consider a facility providing network service charging a price p for every unit of usage and an attachment cost q; we assume there are no usage-dependent costs associated with the facility, and for convenience we consider only non-negative prices. The user population is a continuum labeled by α, with $\alpha \in [0, 1]$. The usage of each user is denoted by x_α. We consider utility functions of the form $U_\alpha = V_\alpha - px_\alpha - q$, where V_α represents the valuation of usage, and assumes users can detach (yielding a utility $U_\alpha = 0$) if prices are too high. The total welfare is given by

$$W = \int d\alpha\, V_\alpha,$$

the total usage $Y = \int d\alpha\, x_\alpha$, and the total revenue by $R = pY + q \int d\alpha$, where the integrals run over all attached users.

The functions V_α take the form $V_\alpha = \lambda_\alpha x_\alpha - x_\alpha^2$. Each attached user sets their

$$x_\alpha = \frac{\lambda_\alpha - p}{2}.$$

We consider three cases for the λ_α: (1) a homoegeneous case (where all users have the same V_α), (2) a heterogeneous case (where users have different V_α) without network externalities, and (3) a heterogeneous case with network externalities (where one user's valuation depends on the number of other attached users).

In the homogeneous case, we set $\lambda_\alpha = 1$ for all α. The total welfare is

$$W = \frac{1 - p^2}{4}$$

as long as

$$0 \le p \le 1 \text{ and } 0 \le q \le \frac{(1-p)^2}{4}$$

(otherwise, all users detach and $W = 0$). Thus, welfare is maximized when $p = 0$, $q \le 1/4$ and each $x_\alpha = 1/2$. Setting $p = 0$ and $q = 1/4$ raises maximal revenue in this case. In this homogeneous case, attachment prices are indeed the optimal way to raise additional revenue.

We now consider a heterogeneous case where $\lambda_\alpha = \alpha$. For a given p and q, all users with $\alpha >$

$$A(p,q) = p + 2\sqrt{q}$$

remain attached. The total welfare is given by

$$W = 1/12(1 - A^3) - \frac{p^2}{4}(1 - A),$$

and the total revenue is given by

$$R = q(1 - A) + \frac{p}{4}(1 - A^2) - \frac{p^2}{2}(1 - A).$$

The total welfare is maximized when $p = 0$ and $q = 0$. The curve of Ramsey prices, the points that maximize W for a fixed R, is given by

$$q = p^2 \text{ for } 0 \le p \le 0.2.$$

The point $p = 0.2$ and $q = 0.04$ maximizes the revenue R. The quadratic nature of the Ramsey curve means that increases in usage prices dominate (over increases in attachment prices) close to the origin.

We introduce network externalities by allowing the constants λ_α to depend on the number of other attached users. Let $\lambda_\alpha = \alpha(1 - A)$ where, as above, A is the critical value of α such that all users with $\alpha > A$ are attached and no users with $\alpha < A$ are attached. Then $A = 1/2[1 - (1 - 8\sqrt{q} - 4p)^{0.5}]$. In addition,

$$W = (1 - A)\left(\frac{(1 - A^3)(1 - A)}{12} - \frac{p^2}{4}\right)$$

and

$$R = (1 - A)\left(\frac{p(1 - A^2)}{4} - \frac{p^2}{2} + q\right).$$

The point $p = 0$ and $q = 0$ maximizes W. The point $p = 0.16$ and $q = 0$ maximizes R. The Ramsey prices fall along the line segment between these two points: $q = 0$ and $0 \le p \le 0.16$. For this model, increasing revenue is best done through increasing usage charges only.

The simple model we considered here is extremely unrealistic and neglects important aspects of the problem such as congestion. However, it does illustrate the basic point that when one has a heterogeneous population containing users who derive marginal benefit from attachment, then raising the attachment prices alone is not necessarily a Ramsey price.

Capacity fillers

The ability to express the capacity in terms of an arbitrary filter provides substantial flexibility for accommodating user needs. In this section we give a concrete example of a sophisticated usage filter. A *token bucket* filter is parameterized by a rate r and a bucket size b. Usage complies with this capacity as long as the cumulative number of units sent in any time interval of length t (for any such t) is bounded above by $rt + b$. This allows bursts of size b, but bounds the long-term average to be no greater than r. A filter might be the composition of three different token buckets, one with $b = 0$, one with $b = \infty$ (in actual practice, this value of b will be chosen to be large but finite, since an infinite-sized b imposes no constraints) and one with an intermediate value of b_i. The values of r associated with the two extremal bucket sizes control very different aspects of the traffic: r_0 describes the allowable peak rate and r_∞ describes the allowable long-term average rate (the actual size of the large but finite b value used in practice determines, along with the associated rate, the time interval over which this long-term average is applied). The intermediate parameters r_i, b_i, describe some intermediate allowable burst rate and size. These different filters should be thought of as constraining the flow on different time scales: the bigger the b the longer the time scale. In general, one might describe a filter as a nonincreasing function $b(r)$; for every r there is a token bucket with parameters r, $b(r)$ applied to the flow. By adjusting these parameters appropriately, one can provision the capacity for various levels of Web-browsing or video consumption.

[43]The seminal paper on the topic is: Deering, S and Cheriton, D 'Multicast routing in datagram internetworks and extended LANs' *ACM Transactions on Computer Systems* 1990 **8** (2) 85–110

[44]Casner, S and Deering, S 'First IETF Internet audiocast' *Computer Communications Review* 1992 **22** 92–97

[45]Shenker, S 'Some fundamental design decisions for the future Internet' *IEEE Journal on Selected Areas in Communications* 1995 **13** 1176–1188

[46]Clark *op cit* Ref 13

[47]These issues are discussed in the references cited in Ref 3, and also in: Clark, D, Zhang, L and Shenker, S 'Supporting real-time applications in an integrated services packet network: architecture and mechanism' *Proceedings of Sigcomm '92* ACM Press, New York (1992) 14–26; Shenker, S 'Service models and pricing policies for an integrated services Internet' in Kahin, B and Keller, J (eds) *Public Access to the Internet* Prentice-Hall, Englewood Cliff (1995) 315–337

[48]A slightly out-of-date overview of this proposed architecture is presented in: Braden, R, Clark, D and Shenker, S 'Integrated services in the Internet architecture: an overview' RFC 1633 (June 1994)

Pergamon

S0308-5961(96)00003-1

Telecommunications Policy, Vol. 20, No. 3, pp. 203–217, 1996

0308-5961/96 $15.00 + 0.00

Service architecture and content provision

The network provider as editor

J MacKie-Mason, S Shenker and H R Varian

There are at least two competing visions for the future National Information Infrastructure. One model is based on the application-blind architecture of the Internet; the other is based on the application-aware architecture of cable television systems and online services. Among application-aware architectures, some are content-aware and some are content-blind. In this paper we examine some consequences of these different network architectures for content provision. Copyright © 1996 Elsevier Science Ltd.

Professor MacKie-Mason may be contacted at the Department of Economics, University of Michigan, Ann Arbor, MI 48109-1220, USA (Tel: +1 313 764 7438; fax: +1 313 763 9181; email: jmm@umich.edu). S Shenker may be contacted at Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304-1314, USA (Tel: +1 415 812 4840; fax: +1 415 812 4471; email: shenker@parc.xerox.com). H R Varian may be contacted at the University of California, Berkeley, USA (email: hal@sims.berkeley.edu)

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. This work was supported by National Science Foundation grant SES-93-20481.

Introduction

The formation of the National Information Infrastructure (NII) is often depicted as a battle between various economic entities such as the cable television companies and the regional Bell operating companies. More recently, the Internet—with its enigmatic origins, anarchic organizaiton and startling growth—has gained widespread visibility and is now seen as a fundamental component of the NII. In addition, the various online services (AOL, Prodigy, CompuServe, MSA, etc), direct broadcast TV, wireless telephony and other emerging technologies will also likely be important parts of the NII.

These various delivery systems display a wide range of characteristics. The physical transmission medium runs the gamut from satellite transmission to coax cables to fiber optics. Some components are broadcast in nature (everyone in the system receives every signal) and others are switched (signals are directed to specific receivers). The corporate entities deploying these component technologies are diverse—from startups to established concerns, and from local monopolies to international competitors—and they face disparate regulatory restraints. These physical, corporate and regulatory issues have been thoroughly discussed in the literature, and we do not address them here. We focus instead on another issue that, to date, has received considerably less attention.

The purpose of these systems is to transport information *content*. We are interested in how the differences between systems affect the offering of content to consumers. We begin with the observation that these systems have radically different *service architectures*. By architecture we are not referring to the actual physical implementation of the network. Rather, we are referring to the nature of the transport service offered. In this paper, we concentrate on one crucial feature of service architecture: *transparency*. We call the entity that transports the bits to

consumers the *network provider*. In some architectures, the network provider is aware of the content of transported bits (eg movie, point-to-point video, music, etc), and in others the content is completely opaque to the network provider.[1] Shenker offers a brief and speculative discussion of the implications of this distinction.[2]

We then ask the question: how does this difference in architecture affect the content provided to consumers? Does the difference in the network provider's awareness affect the selection of existing content that is made available to consumers? These are questions that are getting increasing attention in the press concerned with the strategic developments in competition between the Internet, telephony and cable networks. For example, George Gilder writes that:

> Networks promote choice, choice enhances quality and quality favors morality. Television is culturally erosive because its small range of offerings requires a broad, lowest-common-denominator appeal. Linking to millions of cultural sources, global networks provide a cornucopia of choices, like a Library of Congress at your fingertips. On the Net, as at a giant bookstore, you always get your first choice rather than a lowest-common-denominator choice. A culture of first choices creates a bias toward excellence and virtue.[3]

Also important is whether architecture choice affects the incentives to create *new* content. We address this question in another paper.[4]

In this paper we explore one way in which architecture may affect content provision: through the extent to which the network provider can play an editorial role in selecting the content made available to consumers. In an aware architecture the provider can offer an editorial service; in a blind architecture it cannot. We characterize the different architectures in the next section. Then, in the section that follows, we consider the effects of architecture on the selection of already created content to be offered on the network.

There are other important ways in which service architecture can affect content provision. In a related paper we examine two: through technological and institutional delivery costs that vary across architectures, and through the extent to which architecture permits the network to differentiate transport prices for different goods.[5] We discuss briefly these effects—and some effects on the incentives to create new content—in the fourth section of this paper.

The relationship between architecture and content provision is too complex to yield a simple and definitive answer. Our purpose here is to provide some initial intuition about the effect of architecture on content provision. To that end, we analyze this question in the context of some very simple models; we use these to identify some of the major issues and illustrate them through examples. In particular, throughout our discussion we restrict ourselves to the case of a single network provider that serves customers who choose to connect; our present analysis does not address the more complicated, and realistic, case of multiple coexisting and competing architectures.

Architectures

The architectural distinctions we make here concern the extent to which the network provider distinguishes between the bits it conveys. The Internet, telephony and cable TV occupy very different places along this spectrum; they are representatives of the three basic architectural choices that we describe below.

[1] In the second section we clarify that there are gradations of awareness, from awareness of the applications (eg teleconference vs file transfer) to awareness of the actual content (eg the specific movie being transmitted).

[2] Shenker, S, Clark, D D and Zhang, L 'Services or infrastructure: why we need a network service model' in *Proceedings of Workshop on Community Networking* 1990 145–149

[3] Gilder, G 'Angst and awe on the Internet' *Forbes ASAP* 1995 **4** (December) 132

[4] MacKie-Mason, J K, Shenker, S, and Varian, H R *Network Architecture and Content Provision: An Economic Analysis* Tech rep, University of Michigan, Michigan (1995)

[5] MacKie-Mason, Shenker and Varian *Op cit* Ref 4

Architectural choices

Application-blind networks. One of the Internet's central design principles is that the network provides only bit transportation; it is up to the end-hosts to construct higher-level applications on top of this raw transport mechanism. This architecture has the feature that it need not be modified as new applications arise, because applications are implemented entirely at the end-hosts, and no centralized authority need approve such applications.

We will use the term *application-blind* to refer to architectures where a general interface is made available to end-users, who then implement their applications on top of this interface. For the purposes of our analysis we assume that application-blind networks operate as common carriers. They offer a single, non-discriminatory price for transport and accept any and all traffic at that price.

Network blindness has both advantages and disadvantages. The ability for end-users to develop and implement new applications without network intervention may encourage creativity and experimentation. For example, the Internet has seen a proliferation of applications, such as electronic mail and the WWW, that were not envisioned when the IP protocol was originally designed. However, a blind network provider is unable to provide a gateway, or editorial service, that selects which applications and goods to make available to users. For example, the explosion of content on the Internet has led to an increasing amount of low-quality and irrelevant material that users must search through to find valuable content. A network-cum-editor can filter and certify content, much as the editor of a newspaper does.[6]

Application-aware networks. Telecommunication infrastructures developed by private enterprise such as cable TV and telephony are more tightly coupled to specific applications. The underlying transport function deep within these networks may be general but the interface presented to users is highly restricted. In this type of network the architecture is aware of the type of application that is being used.

For example, telephone networks are designed around voice traffic, and cable TV protocols are designed for video traffic. Similarly, online service providers such as CompuServe and AOL generally know what kinds of applications (voice, data, images, etc) are being transmitted.

By an *application-aware* architecture, we mean that the network service provider can identify the general type of application being invoked (eg email, audio playback, interactive video). This permits some degree of editorial service. For example, AOL permits users to access *Time Magazine*, but for its first several years did not provide access to content on the Internet.

Content-aware networks. Besides being able to determine the general applications customers use, some networks can also monitor and even control the content that is transported over some applications. For example, cable TV can distinguish basic from premium channels, and video-on-demand systems know what movie has been requested. Online services also sometimes know what sort of content is being requested: airline prices, bibliographic references, cartoons, etc. We define a *content-aware* architecture as one in which the network provider can identify the network content.

It is not the case that all *application*-aware networks are also

[6]One might argue that indexing, rating, information filtering and editorial services are emerging on the Internet, but these services merely shift the problem to another level: now the network user must sift through a growing number of filter services to select a preferred one. We are implicitly assuming that the network provider has some advantage in providing an editorial service if the architecture permits it. This advantage might follow from a more credible ability to signal quality, perhaps due to the large fixed investment the network makes in infrastructure.

content-aware. For example, telephony is content-blind, as is electronic mail or WWW browsing.[7] When a network is content aware, the provider can play a more discriminating editorial role.

Architectural implications for content provision

Our fundamental point is that network architecture can have important implications for the nature of information goods made available. Anecdotal evidence is certainly consistent with this view. For instance, the application-blind Internet supports a diverse and rapidly growing set of applications, such as electronic mail, file transfer, teleconferencing and WWW. The application-aware telephone system supports only a narrow range of services (basically fax, low-speed modems and telephony), but provides access to varied content in the form of 900 numbers. Cable television, which is content-aware, offers only one application and rather limited content. Are these differences related to architectural differences? Since the network provider is the entity actually delivering the goods to consumers, the policies of the network provider will affect which goods actually are available to consumers; we investigate the role of architecture in shaping those access decisions.

To focus on the role of architecture, we simplify or eliminate most other relevant factors. For instance, we assume the network provider is a monopoly. This provider is free to maximize profits without competitive (or regulatory) pressures. We also assume that content provision is competitive, with a large number of content providers, and incurs no marginal cost (we can easily incorporate a finite marginal cost into the formalism at the expense of notational simplicity).[8] Therefore, the monopoly network provider can set prices without negotiating with the competitive upstream content providers.

Our modeling of goods and consumers is also quite simple. In most of what follows, we label separate goods by an index i, although in some examples it is more convenient to consider a continuum of goods labeled by $x \in [0, 1]$. These goods all have equivalent bandwidth requirements. We model the consumers as consuming at most one unit of each good, with a reservation value v_i^α for consumer α. Initially, we assume that marginal willingness to pay is independent of the consumption of other goods, so a consumer's satisfaction is merely the sum of the consumed v_i^α less the price paid for the goods and any other non-marginal costs imposed by the network.

In a content-blind but application-aware architecture, different goods denote different applications. In a content-aware architecture, different goods can refer to different content as well as different applications. Given this different definition of goods in the content-aware and application-aware architectures, our analysis need only distinguish between aware and blind architectures. The blind architecture cannot distinguish between goods, and the aware architecture can. This key architectural distinction has several different implications for content selection from already created goods, which we now discuss.

[7]The line between application and content awareness is not always clearly defined. For example, an online service might know that a user is accessing *Time Magazine*, but may not be able to track which articles the user is viewing. Is this application- or content-awareness? In part due to this blurriness, we do not distinguish between the two types of awareness in our present analysis.

[8]With competitive content provision and no marginal content cost, there is nothing to be gained from vertical integration between the network provider and content providers. Nor could *bundling* be used to raise profits. We intentionally sidestep these interesting and important strategy questions in this paper.

Content provision

How do different service architectures affect the provision of already created information content? Our focus in this paper is the opportunity for an aware network to provide an editorial service by controlling the content available. We explore two possible motivations for limiting

content: *clutter costs* and *attention costs*. Clutter cost is an increase in the difficulty of finding or processing information that results from the total number of information goods *available* on the network. Such costs arise at the network level—they are not attributable directly to individual consumer decisions. Attention costs, on the other hand, are a reduction in the value of information goods as the number of goods purchased by a user's increases. These costs can be traced to decisions made by individuals.

We believe that costs of clutter and attention begin to distinguish the economics of information goods from other more traditional fields in economics. Of course, there is nothing novel about externality costs or interdependent demands, which are the formal characterizations we use to specify clutter and attention. However, in the context of information networks, much of our attention in this paper is directed to whether there are too few or too many goods offered. This is fairly unconventional for an economic problem; more choice over available goods is routinely assumed to be unambiguously desirable. Questions of this sort do appear in the economics of advertising, but research in that area has also been somewhat unconventional and plausibly is an early example of the economics of information.[9]

The effect of different types of cost on content provision depends on the network architecture. An aware network—with its greater control—is advantaged when there are significant *network costs*. That is, the selection of which goods to provide is more efficient. There is no such advantage with *user costs* since these can be allocated efficiently by the actions of individual users without network intervention.

The different locus of control also implies a difference in the selection order for content. An aware network controls content selection and thus orders choices by the profits each good generates. (We make our notion of 'ordering' precise below.) In a blind network, with its single transport price p, customers control content selection. Any good with positive demand at price p will be purchased by some consumers, so goods are generally ordered by maximal willingness to pay.

This contrast between ordering by profitability versus maximal willingness to pay has interesting implications for the content diversity on various networks. Consider, for example, two very different kinds of goods: low-value *mass-market* goods, which have low maximal willingness to pay but high total revenue, and high-value *niche* goods, which generate relatively little revenue but have high maximal willingness to pay. Aware architectures will tend to favor low-value mass-market goods, while blind architectures will favor high-value niche goods. We discuss this more fully elsewhere.[10] The bias toward mass-market goods is consistent with what we observe when comparing the offerings of the Internet with cable television.

We now make these observations more concrete by considering how clutter and attention costs affect the provision of goods.

Clutter effects

When there are many information goods or applications available, the *clutter* that results can decrease the value to customers. It becomes harder (and slower) to locate the desired content, or the interface becomes more difficult to use.

Blind networks cannot directly control the variety or quality of content available. They only set a single, uniform transport price that

[9]There is quite a bit of research on the optimal variety of goods when products are differentiated, but that literature is concerned with closely related substitutes. We are investigating markets with very different information goods; indeed, when we model clutter costs we assume that demands for different goods are strictly independent.

[10]MacKie-Mason, Shenker and Varian *op cit* Ref 4

[11]This is not to imply a causal link; there are many differences between these two systems that could explain this observation. For instance, the broadcast nature of cable television would favor mass-market goods, while the switched nature of the Internet does not, although the advent of multicast is changing that somewhat.

acts as a cut-off. Any good or application worth more to some user than the uniform cut-off price will be offered, regardless of the clutter costs imposed on other users. The concern with the resulting clutter in blind architectures is evidenced by the huge popularity of Web indexing services (http://www.yahoo.com/ receives about 1.5 million visits per week) and the demand for them from commercial providers (for example, AOL recently purchased WebCrawler).[12] Readers of Usenet newsgroups and Internet mailing lists are also familiar with clutter costs. The current Usenet feed is about 5 MB of new, mostly unmoderated text material per day.

Content-aware architectures can control the content available to users and play an editorial role; in fact, this is already standard practice in the moderated discussion groups of many online services. Online services also restrict the variety and quantity of content available.[13]

Clutter costs are *network costs*—they depend on the total number of applications or goods available, not the choices made by particular individuals. However, these costs arise not as expenses directly incurred by the network in providing a service, but in reduced user utility as more goods are offered. We model clutter effects by decreasing an individual's utility as more goods are offered.[14] For instance, if the network offers a set of goods G, and a consumer α has purchased a set of G_α goods, then the total valuation by that consumer is $\Sigma_{i\epsilon G\alpha}\, v_i^\alpha - F_c(|G|)$ where F_c is some non-decreasing function. If $\Sigma_{i\epsilon G\alpha}\, v_i^\alpha - F_c(|G|) < 0$ we assume the consumer leaves the network. Let C denote the set of connected consumers, those for whom $0 \leq (\Sigma_{i\epsilon G}(v_i^\alpha - p_i) + - F_c(|G|))$.

A network with a blind architecture sets a uniform transport price p. Let the revenue to a content provider for good i be $R_i(p)$. Then provider i will offer its good if and only if $R_i(p) > 0$. However, adding a good increases clutter costs for all network users: some marginal customers may disconnect. Thus, revenue must be computed over those consumers still attached to the network, $\alpha \in C$. From this we can make the following observation about content provision in a blind network with clutter costs: the selection of goods is ordered by maximal willingness to pay. That is, any good j for which $\max_{\alpha\epsilon C}[v_j^\alpha] > p$ will be offered. Thus, *the users* select the content available, rather than delegating the selection decision to the network.

In the presence of clutter, a network provider with an aware architecture can decide which goods to make available. The problem is to choose this set G and the associated p_i to maximize the total revenue, which is given by the sum $\Sigma_{\alpha\epsilon C}\Sigma_i p_i \delta(v_i^\alpha \geq p_i)$.[15] This maximization problem depends on the details of the distribution of consumer preferences. The striking difference from the blind network is that the goods are ordered by maximal revenue rather than by maximal willingness to pay.[16] This is the natural consequence of the locus of editorial control: the aware net makes a profit-maximizing choice for the network of users as a whole; in a blind network each individual user self-selects the desired content.

The different ordering of content selection can be interpreted as a bias towards mass-market goods in an aware network. Offering a new good has an opportunity cost: the revenues lost from customers who detach due to the increased clutter plus the revenue lost from lowering some prices to keep other customers attached. This opportunity cost must be overcome for a new good to be offered. Consider two new goods with identical value to all users, but which appeal to different

[12]The field of Web indexing services is itself getting quite cluttered now: there are at least a dozen competing services available. To help users choose an index service—so that they may then use the index to help them choose content—we have already seen several comparative studies of index services; see, for example: http://www.zdnet.com/pccomp/features/internet/search/index.html. Also there are several meta-search services (that search multiple search services), such as the All-in-One Search Page, CUSI, Internet Exploration Page, and SavvySearch. That is, we are already seeing editorial meta-services that recommend editorial services!

[13]Interestingly, the online services have recently engaged in a mixed strategy, providing both an editorially controlled service and a gateway to unrestricted Internet content.

[14]See Shiman, D *When E-mail Becomes Junk Mail: The Welfare Implications of the Advancement of Communications Technology* Tech rep, State University of New York, Oswego (1995) for a discussion of similar congestion effects for electronic mail.

[15]The delta function is defined $\delta(z) = 1$ if z is true, 0 otherwise.

[16]The revenue again must be evaluated with respect to those consumers still connected to the network, given the other goods offered. We can formalize the notion of profit ordering by using a necessary condition for offering good j, condition on the set of other goods that are affected at an optimum. The incremental good j is sold to some subset of connected users, yielding revenue $R_j(p_j)$. However, some users detach from the network as a result of the extra clutter. Let D_j denote the set of users who detach, that is, those for whom $F_c(|G_{-j}|) - F_c(|G|) - v_j^\alpha \delta(v_j^\alpha > p_j) > S_{-j}^\alpha$, where G_{-j} is the set of offered goods excluding good j, and S_{-j}^α is the surplus obtained by the user α from goods G_{-j}. Then, we can say that goods are ordered by revenue in the sense that all offered goods satisfy $R_j(p_j) > \Sigma_{\alpha\epsilon Dj}\, \Sigma_{i\epsilon G\alpha}\, p_i$, and all other goods do not.

fractions of the user base. The mass-market good is purchased by f_m users; the niche good by f_n users; and $f_m > f_n$. Then the revenue from the mass-market good with f_m customers will be greater than the niche good revenue, and it will more likely exceed the opportunity cost. Indeed, even if the niche good was more highly valued by users, the aware network favors the mass-market good if the difference in market size is large enough: $f_m p_m > f_n p_n$.[17]

When there are clutter costs a network with an aware architecture offers different goods than does a blind network, but does it handle clutter better? Clutter costs are an *externality*: offering a new good benefits some users but hurts all others. In a blind network, content providers will decide to enter the market as long as there is a single user with $v_j^\alpha > p$; the resulting clutter will tend to reduce total consumer surplus. This is the typical tragedy-of-the-commons phenomena present in many congestion models.[18,19] In general, an aware network provider will not only select different goods to offer, but will also limit the number of goods offered in order to reduce total clutter costs.

This is easy to see in the case when all users who purchase a good value it the same; then the aware network providers charge a price equal to this value for each good. In order to induce the consumer to connect, the provider pays a subscription rebate equal to the user's clutter costs, $F_c(|G|)$. In this way, the network provider extracts exactly $\Sigma_{i \in G\alpha} v_i^\alpha - F_c(|G|)$ from each consumer α, or the total consumer surplus. Since the aware provider is maximizing total surplus, it selects the first-best set of goods—the clutter externality is internalized. More generally, the network will not achieve the first best. However, reductions in the consumer's surplus due to clutter do reduce the profit it can extract, so the aware provider acts as an editor by limiting the number of goods offered in order to reduce total clutter costs.

The clutter cost externality has important implications for interpreting the different content ordering we discussed above. Recall that the aware architecture will favor mass-market over niche goods (appropriately defined). This is not necessarily a bad thing. A niche good, by definition, benefits a small subset of users, while imposing additional clutter cost on all others. On net, total welfare may be lessened by introducing a niche good. In the Appendix we show that, indeed, too many niche goods will be available on a blind network.

We have characterized two effects that service architecture has on content provision when there are clutter costs. First, an aware network offers content based on a profit ordering; content is offered on a blind network according to maximal willingness to pay. Second, the clutter cost is an externality that can be at least partially ameliorated by an aware network provider that acts as an editor or gateway. We now illustrate these aspects of the clutter effect through two numerical examples. A more general treatment is presented in the Appendix.

Example 1: no blind equilibrium. There is a continuum of goods labeled by x, with $x \in [0, 1]$. Each individual consumer α values each of these goods an amount v^α, and these values v^α are uniformly distributed between 0 and 1 throughout the population.

Let p be the transport price of these goods (which are priced the same even in the aware architecture since they are essentially identical). Suppose a network offers x goods; if $v^\alpha > p$ then user α connects and buys all x of the goods, to receive utility

[17] Although surely not due to clutter alone, the reports that about one-quarter of users of online services like AOL detach within a year suggests the number of marginally attached customers is sufficiently large that this bias in selecting new content could be important. See MacKie-Mason, Shenker and Varian *op cit* Ref 4 for a more complete discussion on the mass-market bias in aware networks.

[18] See MacKie-Mason, J K and Varian, H R 'Pricing congestible network resources' *IEEE Journal of Selected Areas in Communications* 1995 **13** (7) (available at URL: ftp://gopher.econ.lsa.umich.edu/pub/Papers/pricing-congestible.ps.Z) for an analysis of congestible resources.

[19] Another aspect, which we have not explored here (in fact, it is precluded by our particular modelling of the clutter effect), is that if different goods experience different amounts of clutter (ie are devalued differently), then goods that are less susceptible to clutter can displace those that are more susceptible to clutter, even if their intrinsic value is less. This is akin to the effect that in a computer network congestion-tolerant traffic can squeeze out congestion-intolerant traffic, even though the congestion-intolerant traffic may intrinsically be much more valuable.

$$U_\alpha = \int_{1-x}^{1}(v^\alpha - p)\mathrm{d}s - F_c(x). \quad (1)$$

We assume that the clutter effect function F_c is of the form: $F_c(z) = \lambda z^2$. By solving for $U_\alpha = 0$ we find that all users with $v^\alpha > p + \lambda x$ are connected, and thus, the fraction of connected users and the total demand for each offered good is $1 - p - \lambda x$. Total profit is calculated as demand price times the number of goods, or $\pi = (1 - p - \lambda x)px$.

An aware network can set both p and x to maximize profit; the maximizing values are $p = \frac{1}{3}$ and $x = \frac{1}{3\lambda}$, yielding a demand of $\frac{1}{3}$, a revenue of $\frac{1}{9}$ and a consumer surplus of $\frac{1}{6}$ per unit good. Goods are offered as long as they yield revenue at least $\frac{1}{9}$; at $x = \frac{1}{3\lambda}$, any additional goods would yield less revenue. Although the example is stylized, in that each good is essentially the same, it illustrates the result that goods are selected based on a profit condition in an aware network.

A network with a blind architecture can set p but not x; content providers will enter the market as long as some customers are willing to buy. For any p, as long as the demand per good $1 - p - \lambda x$ is positive, new goods will enter, ultimately driving the demand to zero. Thus, in the blind architecture no stable equilibrium can have positive demand, revenue, or consumer surplus. Given the set of offered goods, a new good will be purchased (and profit will be positive, so it will be offered) as long as at least one customer values it more than the transport price; that is, if $\max_\alpha v^\alpha > p$. Since the optimal price is

$$p = \frac{1-\lambda x}{2} < 1 \text{ and } \max_\alpha v^\alpha = 1,$$

all goods are offered, until congestion crowds out all users. This illustrates both that goods are selected according to maximal user value in a blind architecture, and that clutter creates a congestion externality.

Example 2: an inefficient blind equilibrium. Consider an example with a continuum of goods labelled by x, with $x \in [0, 1]$. Each individual consumer values each of these goods an amount x (ie the xth good is valued an amount x). The clutter effect function F_c is: $F_c(z) = \frac{1}{2}(z - \frac{1}{2} + \epsilon)_+$. In the blind architecture, all goods have the same price p and all goods with $x \geq p$ are offered. Since $R = p(1-p)$, revenue is maximized at $p = \frac{1}{2}$ and the resulting revenue is $R = \frac{1}{4}$. The consumer surplus is $S = \frac{1}{2}(1-p)^2 - \frac{1}{2}(\frac{1}{2} - p + \epsilon)_+$, which takes on the value

$$\frac{1-4\epsilon}{8}$$

when $p = \frac{1}{2}$. Then, for sufficiently small $\epsilon > 0$, the total welfare and the consumer surplus are increasing in the region $\frac{1}{2} < p < \frac{1}{4} + \epsilon$, even though the blind revenue is maximized at $p = \frac{1}{2}$. This last statement can be verified by noting that the price derivative of the total welfare is $D'(p)(p - F'(D(p))$ where $D(p)$ is the demand function. Thus, the total welfare increases with p if and only if $F'(D(p)) > p$.

In this example the clutter externality in a blind network can be resolved with a traditional Pigovian tax to raise the price by ϵ. As an alternative, an aware network fully internalizes the clutter cost and achieves the first best. Of course, consumers would prefer the blind network because they get to share in the surplus, all of which is extracted by the network provider before the architecture is aware.

Summary. In the Appendix, we present a more general model of content provision with a clutter externality. In that model, a Pigovian tax may lower welfare in a blind architecture, in contrast to Example 2 above. However, the editorial service provided by the aware network against internalizes the clutter problem. We formally show that blind network will offer *too many* goods, and that the first-best involves eliminating some *niche* goods.

In our analysis of content selection with clutter costs we have uncovered the following two principles.

- *Externality*: Content value is reduced as the menu of offerings becomes more cluttered. An aware network provider can serve a beneficial editorial function, increasing the value of the net by limiting the offerings.
- *Content ordering*: Since the aware net provider controls the offerings, content selection will be ordered by profit under an aware architecture and by willingness to pay under a blind architecture.

Suppose that network providers or others develop ways to control clutter cost even in a blind architecture. Even so, an individual user has limited time and attention to devote to different information goals. We now explore how network architecture affects content provision when users experience attention costs.

Attention effects

In the previous subsection we explored the fact that when many goods are purchased by a consumer, his or her satisfaction may be significantly less than the sum of his or her v_i^αs. We assumed that the whole may be less than the sum of the parts due to clutter: the more goods or applications available, the harder it is for the user to find what he/she wants. Another possibility is that users have limited attention, and their enjoyment of a given good is decreased when they are also consuming other goods. There has been considerable recent discussion of the extent to which we are moving from a service economy to an *attention* economy.[20] For example, a subscription to HBO has reduced value if the user also subscribes to additional movie channels. This is a special case of the goods being (imperfect) substitutes for each other. Attention depends on the goods *consumed* whereas clutter results from the goods *offered*.

It might appear that attention effects are a modest variant on clutter effects and not worth the bother of separate analysis. Indeed, we initially conjectured that the same results would hold: we thought content selection would be ordered differently in aware and blind networks, and that the aware architecture would have an advantage because it could limit the attention cost imposed on users.

In fact, both of these conjectures turn out to be wrong. With a little thought, it may be obvious why they are wrong. However, it is worthwhile to explore how attention effects differ from clutter effects. For one thing, demand interactions of this sort surely *are* relevant for an aware net provider selecting content (though not in the way we originally guessed). In addition, the economics of service architecture is a novel topic, and it is instructive to understand how very slight differences in modeling assumptions can lead to quite different results.

No externality. The first thing to notice is that the attention effect

[20]See, for example, Lanham, R A *The Electronic Word: Democracy, Technology, and the Arts* University of Chicago Press, Chicago (1993)

does not create an externality. Attention cost is not imposed on users—they impose it on themselves. Suppose we write individual utility as $\Sigma i \epsilon G_\alpha v_i^\alpha - F_a(|G_\alpha|)$ for a consumer who purchases a set G_α of goods, where $F_a(\cdot)$ is the non-decreasing attention cost of consuming those goods. The $F_a(\cdot)$ term represents interdependence in a single user's utility function; it is unaffected by what goods other users consume. Thus, there is nothing intrinsic about the attention affect that favors the aware architecture over the blind.

We use a simple example with homogeneous users and heterogeneous goods to illustrate the absence of an attention externality in a blind network. Consider a continuum of goods labeled $x \in [0, 1]$, where all consumers agree that x is the value of the good. In a blind architecture, a user will purchase all goods that have value $x > q \geq p$. The cutoff q will generally be greater than the price p because for each good consumed the user pays both p and an incremental attention cost, F_a'. Therefore, surplus is

$$U = \int_q^1 (v - p)dv - F_a(1 - q). \tag{2}$$

We suppose that the attention cost function is given by

$$F_a(z) = \tfrac{1}{2}(z - \tfrac{1}{8})_+.$$

The consumer chooses to purchase all goods with greater value than $q = \frac{1}{2} + p$, where the $\frac{1}{2}$ is simply the marginal attention cost. Substituting back into the surplus function and taking the derivative with respect to p we find that

$$\frac{\partial U}{\partial p} = \begin{cases} -\tfrac{1}{2} + p, \text{ if } p \leq \tfrac{3}{8} \\ -1 + p, \text{ if } \tfrac{3}{8} < p \leq 1. \end{cases} \tag{3}$$

For all feasible values of the transport price, surplus is decreasing in price. There is no externality, and reducing the set of offered goods (by raising price) can never make consumers better off.[21]

Content selection by value or profit? The attention effect is a function of consumed goods, not offered goods, and users choose the goods consumed. As a result, it turns out that in both architectures the goods are generally ordered by maximal willingness to pay rather than total revenue. Why is this result different from the ordering by revenue for the aware architecture with clutter effects? The clutter cost is a per good cost borne by the network as a whole; it lowers the total consumer surplus available for extraction. This network-wide cost from offering an incremental good must be recovered. Attention cost, on the other hand, is a per good cost borne by individual users. Since there is no network-wide cost to be recovered, the network offers goods based on individual consumer valuations, not total revenue.

As before, the result is trivial for a blind network: the net provider merely sets the uniform transport price p, and all goods that some consumer values more than p are offered (consumer valuation is net of the marginal attention cost induced by consuming the good). For an aware network, the result is also straightforward. The net provider should offer all goods for which $\max_\alpha[v_i^\alpha - F'(|G_\alpha|)] > 0$.[22] To see why this is so, suppose not for good j: if the network adds j to its offerings, then at least one user would experience higher utility by purchasing it, and there are no external effects on other users, so no one detaches. Thus the net could charge a $p_j > 0$ and increase its profits.

[21] It would never make sense to eliminate a different set of goods, because all consumers agree on the value ordering of the goods. Thus, raising price eliminates those goods that are valued least by all consumers.

[22] We abuse the notation slightly by referring to the incremental change in attention costs as the derivative F_a'. In fact, the argument changes by integer values, and the incremental change is a first difference.

It may be worth noting that ordering by willingness to pay is equivalent to ordering by profit in this case. The point is not that the network does not order by profit—of course, it must if it is maximizing profits—but rather that with only attention costs are the two orderings identical.

Architecture choice and social welfare

We have characterized the effects that choice of service architecture have on content provision. What are the welfare consequences of architecture choice? Given our focus on the selection of goods offered, it is natural to first compare how many goods are made available under different architectures. However, since different goods will be valued differently, we are also interested in how content selection affects the total surplus obtained: that is, the sum of profits and consumers' surplus.

Consider first a network with attention costs. The aware provider can charge a different price for each good or application; the blind provider chooses only one price. As a result, profits will be higher, and more goods will be offered in an aware network. The aware network number of goods is greater because all goods are offered for which $\max_\alpha[v_i^\alpha - F_\alpha'] > 0$, whereas in the blind net only goods for which $\max_\alpha[v_i^\alpha - F_\alpha'] > p > 0$ are offered.

How do the number of goods provided by aware or blind monopoly network providers compare to the first-best welfare optimum? It is easy to determine the welfare-maximizing outcome when there are only attention costs. Each user orders all of the goods from highest v_i^α to lowest. The user then adds the goods to his/her consumed set in order until $v_j^\alpha < F'(j)$. The union of the goods desired by all users is the optimal offered set. That is, all goods for which $\max_\alpha [v_j^\alpha - F'(|G_\alpha|)] > 0$ should be offered. This is the set of goods offered by the aware network, so the aware provider makes the optimal selection, whereas a blind network offers too few goods.

Even though the aware architecture offers the optimal number of goods when there are attention costs, both architectures achieve less than the optimal level of social welfare. For the blind architecture, the result follows directly from offering too few goods. In an aware network, the provider chooses prices $p_i > 0$. Now consumers order the goods by $v_i^\alpha - p_i$. Thus, the order in which the consumer selects goods is distorted. Further, the consumer purchases goods only until $v_j^\alpha - p < F_a'$, so an individual consumer takes fewer goods than the social optimum ($v_j^\alpha < F_a'$)—that is, the right set of goods is offered, but too few consumers buy them. This result contrasts strongly with the result for a network with clutter costs: in that case the provider used its editorial opportunity to offer *fewer* goods under an aware architecture.

Although network pricing lowers social welfare when there are attention costs, there is no market failure as there was with the clutter effect. The result is simply the usual monopoly result: the net provider restricts the output in order to raise prices above marginal cost and earn supracompetitive profits. The editorial role in an aware network is not an intrinsic advantage when there are attention costs; private consumers internalize the problem and solve it themselves.

Both architectures offer suboptimal social welfare when there are attention effects. Can we say which does better? Unfortunately, no. Though more different goods are offered in an aware net, $p_i < p$ for

[23]That is, more different goods are offered, but less is consumed of some of the goods.

some i, $p_j > p$ for some j, and the net effect on social welfare is ambiguous.

What happens to the number of goods offered and social surplus when there are clutter costs rather than attention costs? We showed in the section on clutter effects that the blind architecture offers fewer goods than an aware network. This provides another sharp contrast between clutter and attention effects, since a blind network offers more goods in the latter case.

How does an aware network compare to the social optimum when there are clutter costs? We showed earlier that when all purchasers of good i have the same valuation for that good, $v_i^\alpha = v_i^\beta$, the aware provider could achieve the first best by charging $p_i = v_i$ for the goods and then giving a subscription rebate to users to cover their clutter costs (assuming that everyone experiences the same clutter costs). Now suppose there is some variation across users in the clutter cost function. Then some users will get positive surplus from the original subscription rebate, while others will want to detach. To retain some of the surplus from those users who want to detach, the network will exclude some low profit goods to reduce clutter costs. Thus, we expect to see too few goods on an aware network with clutter costs, consistent with the usual monopoly result. We cannot be sure whether the blind net offers too few or too many goods relative to the optimum because the monopoly profit incentive and the clutter externality effect work in opposite directions.

Unfortunately, the effect of architecture choice on total surplus is ambiguous when there are clutter costs, as it is for attention costs. In general, the welfare will be below the optimum with both architectures when the network provider is a monopolist. However, either the blind or the aware architecture could surpass the other, depending on consumer preferences.

Price differentiation and provider costs

In another paper we have explored ways in which the difference in service architecture can affect both content provision and the incentives to create new content.[24] We summarize some of the main results here to flesh out a richer view of our ideas on the economics of service architecture.

Suppose there are no clutter or attention costs, or other consumer disutilities. Then the main feature of the difference in architecture is the fact that an aware network can charge a different price for transporting different goods or applications, while a blind network is limited to a single price. We also add one element of provider cost ignored in the present paper: there is a fixed cost per good offered that a network with an aware architecture must pay. We refer to this as a gateway or liability cost. For example, the aware architecture might have to reprogram to add a gateway to deliver a new application. Alternatively, under current and emerging US law (at least), an aware network provider may be liable for certain types of content (eg libel or obscenity) that it transports.

When we compare the aware and blind architectures under these conditions, we find that potentially severe inefficiencies and political economy conflicts can arise. For example, if gateway costs are relatively low, the monopolist net provider will generally prefer an aware

[24]MacKie-Mason, Shenker and Varian, *op cit* Ref 4

architecture (since it can imitate blind price as one option, or do something better). But consumers will often (not always) prefer the blind architecture, because they can retain more surplus. Total welfare could be better under either architecture, depending on the specifics.

For example, suppose most consumer variation in v_i^α is in i (that is, 'within' variation with most users having the same preferences but each having widely varying valuations across goods). Then the aware architecture has higher total welfare (and network profit), but consumers prefer the blind architecture. If, on the other hand, most variation is 'between'—different valuations for the same good by different consumers—then both consumer surplus and total welfare can be higher under a blind architecture.

When we examine the creation of content, we find an interesting problem of commitment. The aware network can extract all of the surplus from a new information good, while the blind architecture leaves more surplus as a reward—and thus incentive—for the creator. However, as above, we note that the aware network can mimic the pricing policies of the blind network. Therefore, it might seem an aware network should be able to induce at least as much creative effort as a blind network. However, the mimicry strategy is not compelling here, because the problem is dynamic: the aware network cannot credibly commit that it will not expropriate the surplus in a future period, after the investment in creation activity is sunk. By choosing a blind architecture, then a network may be making a credible commitment to leave incremental surplus in the hands of creators and thus it may induce a higher steady-state level of creative activity.

There is another obvious difference that is likely to affect content *creation*: in an aware network, a single firm (the network provider) effectively decides whether to invest in creating new content, while multiple firms each make independent decisions in a blind network. We expect that in a blind network there will be more experimentation with content creation and collectively more risk-taking (not because of differences in risk aversion, but because of differences in beliefs about the likely success of projects).

Discussion

One key distinction between various competing visions for the NII is the extent to which the network provider is aware of the content of the bits it is conveying to consumers. Aware architectures are aware of the content, while blind architectures are not. Our focus, in this paper, is the impact this architectural distinction has on the provision of content.

The most striking difference between networks with aware and blind architectures is that the selection of offered goods proceeds by maximal profit in aware networks, whereas goods are ordered by the maximal willingness to pay among users in a blind network. We noted that in an aware network this ordering favors mass-market over niche goods. These results apply, formally, when there are either clutter or attention costs. However, when the costs take the *attention* form (and if there are no other network-wide costs from offering an incremental good), then the profit and willingness to pay orderings coincide.

Even though the selection orderings coincide when there are only attention costs, the *number of goods* offered under the different architectures do not: an aware network will offer more goods when

there are attention costs. This occurs because the aware network provider can differentiate between goods and thus will offer some low value goods that do not pass the uniform price threshold on a blind network.

In striking contrast, when the costs are experienced as clutter, a network provider with an aware architecture offers *fewer* goods than a blind network. This is the consequence of our third major finding on content provision: clutter costs are an *externality*, but they can be internalized if the architecture is aware. Therefore, in a blind network with clutter, too many goods are offered; the aware network provider exercises its editorial capability to reduce the number of goods offered.

We have studied a very simplified and stylized model of network service architecture and content provision. We plan to explore the effects of service architecture in richer settings. For example, we would expect more than one network provider to compete. Then it is easy to imagine that some customers who find clutter very costly will choose to subscribe to an aware network that controls clutter, at the cost of receiving mostly mass-market goods. Other networks might cater to customers with niche tastes but sufficiently high values for these niche goods to overcome the resulting clutter costs.

However, once we begin to think about multiple networks with multiple architectures and different menus of available content, we have to take seriously another feature of information economies: positive network externalities. That is, the value of belonging to a network for many consumers tends to increase the more other users who are connected. For example, email is much more valuable if it can be exchanged with anyone, not just to subscribers on one of several competing networks. Therefore, we expect to see a growing demand for interoperability. Interoperability between multiple proprietary networks with multiple incompatible architectures is costly and difficult: interoperability seems to favor the blind architecture. Thus, rather than multiple competing network 'clubs', we might eventually see a unified internetwork with a blind architecture, on which customer types who suffer high clutter costs pay for editorial services provided by competing sellers over the blind network. Such editorial services may not be as cost-effective or able to build a reputation as editorial control provided by a network provider, but they may be the second-best compromise that results from the value of interoperability. Our casual observation suggests this vision is consistent with the recent movement of proprietary online services towards providing a gateway to blind Internet servicers, while attempting to differentiate their products through the quality of their competing editorial services.

Appendix

Clutter as an externality

One key aspect of the clutter effect is that the offering of a good affects all consumers negatively (actually, it only affects those consumers who remain connected), but many only offer some user benefits. This negative effect can be controlled by the aware architecture, but not in the blind architecture.

Suppose there are many possible goods, and that each good appeals to an entirely separate group of like customers. That is, for each good i, there are f_i potential customers, each of whom value the good at v_i, and no consumer wants more than one good. Order the index numbers for the goods so that $v_i > v_{i+1}$. Everyone bears the same clutter costs, which depend solely on the number of goods offered: $F(N)$.

In a blind architecture, the network provider can set a single transport price. Customers will participate in the network only if their surplus is positive, so we obtain an individual rationality (IR) constraint for each group of the form $v_i - F(N) - p > 0$, given N. Given the ordering of the goods, $N = \arg\min_i[v_i - F(N)]$, so the IR constraint can be binding for the least valued good, at most. A profit-maximizing network will thus set $p = v_N - F(N)$, condition on its choice of N. That is, the price will be set to extract all of the surplus of the group with the lowest surplus. The network then chooses N to maximize its profits where the profit function is

$$\pi(N) = [v_N - F(N)]\Sigma_{i=1}^{N} f_i.$$

Now we can consider some welfare consequences of the clutter externality.

In general, we would expect that because clutter imposes an externality, there will be *too many* goods in equilibrium. The first way to investigate this conjecture is to ask whether welfare increases if the prices were increased. Consumer surplus at price $p(N-j)$ is

$$\begin{aligned} S(\mathrm{N}-\mathrm{j}) &= \sum_{i=1}^{N-j} f_i[\mathrm{v_i} - \mathrm{F}(N-j) \\ &\quad - p(N-j)] \\ &= \sum_{i=1}^{N-j-1} f_i[v_i - v_{N-j}], \quad \text{(A1)} \end{aligned}$$

where the simplification results from substituting in

$$p(N-j) = v_{N-j} - F(N-j).$$

The change in consumer's surplus from raising price from the profit-maximizing level $p(N)$ to any higher price that is profit-maximizing for the smaller number of offered goods is:

$$\begin{aligned} S(N-j) - S(N) &= \sum_{i=1}^{N-j-1} f_i[v_i - v_{N-j}] \\ &\quad - \sum_{i=1}^{N-1} f_i[v_i - v_N] \\ &= \sum_{i=1}^{N-j-1} f_i[v_{n-j} - v_N] \\ &\quad - \sum_{i=N-j}^{N-1} f_i[v_i - v_N] < 0. \quad \text{(A2)} \end{aligned}$$

Thus, forcing the network provider to charge a higher price (say, by imposing a tax) discouraging some low value goods from being offered will in fact lower consumption surplus, despite the clutter externality. Since we already know that a higher price will also lower profits, it follows immediately that a higher price will also lower total welfare.

This does not mean that the optimal set of goods is being offered, however. Suppose it were possible to pick and choose which goods would be offered, which is precisely what an aware network can do. How would total welfare by changed by eliminating a single good?. Let G_{-j} refer to the set of goods when the jth good is removed from the profit-maximizing set. The change in total welfare from removing the good is

$$\begin{aligned} &W(|G_{-j}|) - W(N) \\ &\quad = \sum_{i \epsilon G_{-j}} f_i[v_i - F(N-1)] \\ &\quad - \sum_{i=1}^{N} f_i[v_i - F(N)] \\ &\quad = -f_j[v_j - F(N)] \\ &\quad + [F(N) - F(N-1)] \sum_{i \epsilon G_{-j}} f_i \quad \text{(A3)} \end{aligned}$$

The first term reflects the loss of surplus from excluding the jth group of customers; the summation reflects the reduction in clutter costs that offering the jth good imposes on all other customers. When the clutter savings are large relative to the surplus from the jth good, welfare would increase by excluding the good. This will tend to be the case for niche goods (small f_j relative to $\Sigma_{i \epsilon \hat{G}} f_i$), and, of course, when the marginal clutter cost

$$F(N) - \mathrm{F}(N-1)$$

is large.

Does an aware network necessarily deal better with the congestion problems? Yes: aware profits are maximized by extracting the full consumer's surplus with $p_j = v_j - F(N)$, so the first term in the expression is the negative revenue from good j, and the aware network drops goods with the lowest revenue until this foregone revenue exceeds the additional revenue (surplus) that can be extracted from other users as clutter decreases. Thus, the aware network completely solves the externality problem, and the welfare-maximizing set of goods is offered.

As we have seen, when there are clutter effects an aware network will order content selection by maximal profit, while a blind network will order by maximal willingness to pay, just as with the liability/gateway effect. However, with a clutter externality, too many goods will be offered in a blind architecture, particularly too many niche goods. Forcing the network provider to raise its transport price may not solve the clutter problem and, in fact, may reduce both consumer surplus and total welfare. Adopting an aware architecture will internalize the clutter externality and, in the special case in which the network can extract all surplus, will even result in the socially optimal set of goods.

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 Erllbaum 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 $\partial U^{ij}/\partial x_{ij}\ 0 \geq 0$ 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, $(\partial U^{ij}/\partial Q^{ii})(\partial Q^{ii}/\partial x_{ij})$ 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,

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

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 $X_i^{Off} = \Sigma_{k \in M, i \neq k} N_{ik}x_{ik}$ is the internet traffic. The internet traffic that either terminates on network 'i' or is pure transit traffic is given by $X_i^{In} = \Sigma_{kj \in z, k \neq i, i \in R(kj)} N_{kj}x_{kj}$. 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:

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

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:

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

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:

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

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:

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

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:

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

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:

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

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

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.

Telecommunications

A Bridge to the 21st Century

Edited by **M. Jussawalla**

Global changes in policy and technology in the telecommunications industry in the 1990s are described and analysed in this volume, showing how this industry creates a bridge for society's transition to the next century. Dynamic innovations in technology are encouraging a relatively free world market and a Global Information Infrastructure for use by developed and developing economies. The volume discusses the challenges posed by these innovations for closing the gap between the information-rich and information-poor. Lessons are included for corporate and individual users as they prepare for the Global Information Infrastructure. Societal impacts of new networks, multimedia, cellular communications, equipment standards, telemedia, digital cash, the Internet, and innovative satellite systems are explored in detail with a view to future developments.

Contents:

©1995 380 pages
Hardbound
Price: Dfl. 210.00 (US$131.25)
ISBN 0-444-82325-5

NORTH-HOLLAND
An Imprint of Elsevier Science

To order please contact one of the addresses below:

Amsterdam
Elsevier Science
Customer Service Department
P.O. Box 211
1000 AE Amsterdam
The Netherlands
Tel.: +31 (20) 485 3757
Fax: +31 (20) 485 3432
E-mail: nlinfo-f@elsevier.nl

New York
Elsevier Science
Customer Service Department
P.O. Box 945
New York, NY 10159-0945
Tel.: +1 (212) 633 3750
Fax: +1 (212) 633 3764
E-mail: usinfo-f@elsevier.com

Tokyo
Elsevier Science
Customer Service Department
20-12 Yushima 3-chome
Bunkyo-ku, Tokyo 113
Japan
Tel.: +81 (3) 3836 0810
Fax: +81 (3) 3839 4344
E-mail:
forinfo-kyf040305@niftyserve.or.jp

Dutch Guilder (Dfl.) price quoted applies worldwide. US$ price quoted may be subject to exchange rate fluctuations.

Beyond Competition: The Future of Telecommunications

Edited by **D.M. Lamberton**

The convergence of telecommunications, mass media and computer technologies has brought spectacular developments of ubiquitous intelligent interconnected systems. In the course of these evolutionary changes, debate and policy has swung again towards privatization, deregulation and increased reliance upon competition. Nevertheless, the underlying and powerful role of new information continues to bring so much restructuring and organizational change, that a reassessment of ideas about competition in this dynamic context, is essential. The aim of this volume is to provide an update of research and policy debates in this important field. An international perspective is provided with contributions from academic, business and governmental communities. The volume will be invaluable to researchers in telecommunications and information activities; decision-makers in industry, government and regulatory fields; consultants; and information service providers.

©1995 434 pages
Hardbound
Price: Dfl. 225.00 (US$140.75)
ISBN 0-444-82252-6

NORTH-HOLLAND
An Imprint of Elsevier Science

To order please contact one of the addresses below:

Amsterdam
Elsevier Science
Customer Service Department
P.O. Box 211
1000 AE Amsterdam
The Netherlands
Tel.: +31 (20) 485 3757
Fax: +31 (20) 485 3432
E-mail: nlinfo-f@elsevier.nl

New York
Elsevier Science
Customer Service Department
P.O. Box 945
New York, NY 10159-0945
Tel.: +1 (212) 633 3750
Fax: +1 (212) 633 3764
E-mail: usinfo-f@elsevier.com

Tokyo
Elsevier Science
Customer Service Department
20-12 Yushima 3-chome
Bunkyo-ku, Tokyo 113
Japan
Tel.: +81 (3) 3836 0810
Fax: +81 (3) 3839 4344
E-mail:
forinfo-kyf040305@niftyserve.or.jp

Dutch Guilder (Dfl.) price quoted applies worldwide. US$ price quoted may be subject to exchange rate fluctuations.

You Benefit With Disk Submission

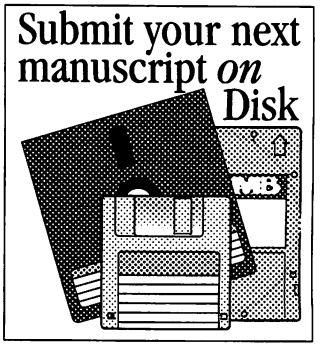
Elsevier Science encourages the submission of articles on disk. We want Authors to provide us with their final manuscripts directly from their computer or word processing systems, without the need to follow complicated instructions. If you submit electronic manuscripts, you will notice the benefits:

- **A speedier publication process**
- **A clearer set of proofs, with fewer errors**

  Providing electronic manuscripts will improve the delivery of proofs - delivery will be faster and more reliable. Also, without needing to rekey the text the possibility of introducing errors will be avoided
- **Inclusion in an electronic archive**

  Elsevier Science is committed to developing and maintaining electronic archives for all of its journals. This will provide the flexibility to take advantage of electronic media to disseminate the information published in our journals, including your article.

There are some basic points to be kept in mind and we do have certain preferences. However, with Elsevier's expertise and facilities it does not really matter on which computer or wordprocessing system your manuscript has been prepared.



Basic Points to Help Us

Delivery of Electronic Files
Disks must be clearly marked with the following information:

- **Operating system**
- **Disk format (e.g.DS/DD)**
- **Word Processor used, including version number (users of see later)**
- **Authors' names**
- **Short title of article**

Three printed copies of the final version of the manuscript should be submitted with the disk to the Journal Editor.
In the event of differences between disk and hardcopy, the hardcopy will be considered as the definitive version.

Preparing Electronic Text Files

Please follow the general instructions on style and arrangement and, inparticular, the reference style as given in Notes for Authors for the book or journal concerned. If you are contributing to a multi-author work, please consult the publisher or earlier volume for style conventions with respect to headings, reference citation, etc.

LaTeX/TeX

Authors wishing to submit their article as a LaTeX/TeX file should note the following:
Authors should ideally use the "Article" style or the Elsevier LaTeX package which is available via anonymous FTP from CTAN centres.
(Further details on TeX can be obtained from Martin Key or Sebastian Rahtz)

Host names:	**CTAN directory:**
ftp.dante.de	/tex-archive/macros/latex/contrib/supported/elsevier
ftp.tex.ac.uk	/tex-archive/macros/latex/contrib/supported/elsevier
ftp.shsu.edu	/tex-archive/macros/latex/contrib/supported/elsevier

Authors should not add their own macros

Preparing Electronic Graphic Files

Illustrations should be produced in the Macintosh environment if possible using the following software packages:

Adobe Illustrator, Aldus Freehand, Cricket Graph, Macdraw, Chemdraw, Corel Draw for PC

However we will accept any of the popular drawing programs for the Macintosh and PC.
Artwork should be drawn for finished size using a Times or Helvetica typeface at a final size of 8pt type with appropriate lineweights.
Please indicate format, operating system, program and version number of the software used. If possible also print a directory of filenames.
Scanned artwork should be saved to Tiff format for both line and halftone and scanned at a suggested setting of 400 dpi for half-tones and 1000 dpi for linework. If it is necessary to compress the scans please indicate the software used. It is essential that a hard copy print of the scans be included. Illustrations should be logically named and saved as individual files to 3.5" disk or a SyQuest cartridge 44Mbyte or 88Mbyte. If 3.5" disks are not available to you, 5.25" disks are acceptable. Please send a laser print of the artwork with the electronic file. When submitting electronic colour images please indicate the file format and program used (including compression software). Include a 4 colour machine or cromalin proof and check that all the separations (if provided) are colour identified.

If you require any further information please contact the following:

Elsevier Science Ltd
The Boulevard
Langford Lane
Kidlington
Oxford
OX5 1GB
UK

Martin Key/Sebastian Rahtz (Text)
Tel: 44 1865 843550
Fax: 44 1865 843905
Email: m.key@elsevier.co.uk
s.rahtz@elsevier.co.uk

Phil Halsey (Graphics)
Tel: 44 1865 843305
Fax: 44 1865 843921
Email:p.halsey@elsevier.co.uk.

Elsevier Science Inc
655 Avenue of the Americas
New York
NY 10010-5107
USA

Tom Lewis Flood
(Text/Graphics)
Tel: 212 633-3855
Fax: 212 633-3658
Email:t.lewisflood@elsevier.com

ELSEVIER SCIENCE

CO611

Telecommunications Policy

Notes for authors

Submissions

Three copies (the original and two copies) must be submitted to the Editor. Articles should be 4000–6000 words long and should refer principally to the political, economic and social aspects of telecommunications. Comments, reports or rejoinders to articles should be much shorter, usually 1000–3000 words.

Contributions are normally received with the understanding that their contents are original, unpublished material and are not being submitted for publication elsewhere. Translated material, which has not been published in English, will also be considered.

The Editors reserve the right to edit or otherwise alter contributions, but authors will receive proofs for approval before publication.

Presentation

Manuscripts must be typed in journal style, double-spaced (abstract and references/notes should be triple-spaced) on one side only of International Standard Size A4 paper (or the nearest size available), with a left-hand margin of 40 mm.

Manuscripts should be arranged in the following order of presentation. *First sheet:* title, subtitle (if desired), author's name, affiliation, full postal address and telephone and fax numbers. Respective affiliations and addresses of co-authors should be clearly indicated. *Second sheet:* a self-contained abstract of 100 words; acknowledgements (if any); article title abbreviated appropriately for use as a running headline. *Subsequent sheets:* main body of text; appendixes; tables (on separate sheets); references and notes (numbered consecutively); captions to illustrations (on a separate sheet); illustrations. Each sheet must carry the abbreviated title of the article and the journal name.

The text should be organized under appropriate section headings, which, ideally, should not be more than 800 words apart. All headings should be placed on the left-hand side of the text, with a double line space above and below.

All measurements should be given in metric units.

Authors are urged to write as concisely as possible, but not at the expense of clarity. Descriptive or explanatory passages, necessary as information but which tend to break up the flow of text, should be put into notes or appendixes.

Text preparation on disk

The publisher encourages submissions to the journal on disk. The electronic version on disk should be sent with the final accepted version of the paper to the the Editor. **The hard copy and electronic files must match exactly.** Please contact the editorial offices for full guidelines on disk submission.

References and notes

These must be indicated in the text by superior arabic numerals which run consecutively through the paper. The author/date (Harvard) system should not be used. References and notes should be grouped together in a section at the end of the text in numerical order and should be triple-spaced. References should conform to current journal style.[1]

[1] *For journals:* Witte, E, and Dowling, M 'Value-added services: regulation and reality in the Federal Republic of Germany' *Telecommunications Policy* 1991 **15** (5) 437–452
For books: Garnaut, R *Australia and the North-east Asian Ascendancy* Australian Government Publishing Service, Canberra (1989)

Tables

Tables should be numbered consecutively in arabic numerals and given a suitable caption. Notes and references within tables should be included with the tables, separately from the main text. Notes should be referred to by superscript letters. All table columns should have an explanatory heading. Tables should not repeat data available elsewhere in the article, eg in an illustration.

Illustrations

All graphs, diagrams and other drawings should be referred to as Figures, which should be numbered consecutively in arabic numerals and placed on separate sheets at the end of the manuscript. Their position should be indicated in the text. All illustrations must have captions, which should be typed on a separate sheet.

Illustrations should be provided in a form suitable for reproduction without retouching: that is, they should be camera-ready. Three copies of the illustrations should be provided: the original, a clean photocopy, and a photocopy with labels marked up as appropriate, in black ink. Illustrations should permit reduction, with lines drawn proportionally thicker and symbols larger than required in the printed version.

Authors should minimize the amount of descriptive matter on graphs or drawings, and refer to curves, points, etc, by their symbols. Descriptive matter should be placed in the caption or a separate note. Scale grids should not be used in the graphs, unless required for actual measurements.

Copyright

All authors must sign the 'Transfer of Copyright' agreement before the article can be published. This transfer agreement enables Elsevier Science Ltd to protect the copyrighted material for the authors, but does not relinquish the author's proprietary rights. The copyright transfer covers the exclusive rights to reproduce and distribute the article, including reprints, photographic reproductions, microform or any other reproductions of similar nature and translations, and includes the right to adapt the article for use in conjunction with computer systems and programs, including reproduction or publication in machine-readable form and incorporation in retrieval systems. Authors are responsible for obtaining from the copyright holder permission to reproduce any figures for which copyright exists.

Proofs

Authors are responsible for ensuring that all manuscripts (whether original or revised) are accurately typed before final submission. Manuscripts will be returned to the author with a set of instructions if they are not submitted according to style. One set of proofs will be sent to the authors before publication, which should be returned **promptly** (by Express Air Mail if outside the UK). The publishers reserve the right to charge for any changes made at the proof stage (other than printer's errors) since the insertion or deletion of a single word may necessitate substantial consequential changes.

Offprints

Fifty offprints of each paper will be provided free of charge to the first-named author of main articles. Further offprints, in minimum quantities of 50, can be purchased from the publisher.

TELECOMMUNICATIONS POLICY

SPECIAL ISSUE

LESSONS FROM THE INTERNET

Volume 20 Number 3 April 1996

Papers

This journal is abstracted and/or indexed in: *Computer Control Abstracts; Electrical & Electronics Abstracts; Physics Abstracts; Communication Abstracts; PAIS Bulletin; Sociological Abstracts; Current Contents/Social & Behavioural Sciences; Current Contents/Engineering, Technical & Applied Sciences; Social Sciences Citation Index; ABI/Inform; Engineering Index Energy Abstracts; Engineering Index Monthly; Engineering Information Inc of New Jersey (Electronics and Communications Abstracts).*

Pergamon

0308-5961(1996)20:3;1-L

3083