OXFORD REVIEW OF ECONOMIC POLICY, VOL. 17, NO. 2

THE ECONOMICS OF THE INTERNET: INFRASTRUCTURE AND REGULATION

MARTIN CAVE
Brunel University
ROBIN MASON
University of Southampton and CEPR[1]

The purpose of this article is to inform users, regulators, and economists about the basic economics of the Internet, focusing on regulation of its infrastructure. It defines the Internet and describes its development and organization. It analyses the regulatory and competition issues associated with conveyance on the Internet. It then discusses three current puzzles of economic interest. While the article reaches several conclusions, the overall message is that much more work is needed in this area.

I. INTRODUCTION

The Internet is a hot topic. The (mis)fortunes of dot-com firms command headline space in newspapers and television reports. Governments have well-publicized policies about extending Internet access to create 'knowledge-based economies'. Grandparents keep in touch with their families by e-mail.

Yet little is known about the Internet. Most people are hard put to say exactly what it is. Regulators and governments know that the Internet is important; but they are not quite sure whether it requires their attention and, if so, how. Economists have, on the whole, not devoted too much effort to the area.

The purpose of this article is to inform users, regulators, and economists about the basic economics of the Internet, focusing on regulation of its infrastructure. Sections II and III provide information about the development of the Internet and its operation and organization. Section IV describes the major regulation and competition issues. Section V describes some current dilemmas in the policy field.

The article reaches several conclusions; they are summarized in section VI. An important overall

[1] We are grateful for helpful comments from Peter Culham and the editors.

message is that much more work is needed in this area to answer all of the outstanding questions. The Internet provides a fertile research area for theoretical and applied researchers in economics—as well as in other disciplines.

II. DEFINITION AND BACKGROUND OF THE INTERNET

The most frequently asked question about the Internet is: what is the Internet? It is surprisingly hard to get a precise answer. The usual definition—'a global network of networks'—is suggestive but hardly exact. A more detailed answer is that the Internet is a worldwide network of computers that use certain protocols for data transmission and exchange. The definitive statement was given on 24 October 1995, when the Federal Networking Council (FNC) unanimously passed a resolution defining the Internet as:

> the global information system that: (i) is logically linked together by a globally unique address space based on the Internet Protocol (IP) or its subsequent extensions/follow-ons; (ii) is able to support communications using the Transmission Control Protocol/Internet Protocol (TCP/IP) suite or its subsequent extensions/follow-ons, and/or other IP-compatible protocols; and (iii) provides, uses or makes accessible, either publicly or privately, high level services layered on the communications and related infrastructure described herein.

For most people, that is unlikely to clear up the mystery. In order to understand the statement, it is helpful to review the history of the Internet, to see how it developed from a connection between two computers in Los Angeles and Stanford to its current position (as of mid-2000) with 93m computers and 407m users. (All estimates of the size of the Internet are unreliable; these were taken from http://www.nua.ie/surveys/how_many_online/index.html and http://www.zakon.org/robert/internet/timeline/.)

The driving force behind the Internet is network externalities—the fact that the value of a set of computers increases with the number of computers that are interconnected. The value of connectivity arises from several sources. Most directly, there are benefits to each individual to being able to communicate with others; the more users are on the network, the greater the total benefit. For example, suppose that each individual gains a benefit of 1 from being able to communicate with any other individual; and suppose that there are N individuals on the network. Then the total value of the network is the number of pairings $N(N-1)$, which is close to N^2 when N is large. This square relationship between the number of members of a network and the value of the network is known as Metcalfe's law. There are also indirect benefits associated with a large network. The more members of the network, the more likely it is that new services will be offered over it. (Think about the increase in the number and range of programmes on television over the last 50 years, as the number of television owners has risen.) In short, networks are more valuable if there are more people using them. See Katz and Shapiro (1985) and Farrell and Saloner (1985, 1986) for seminal analyses of network externalities.

The Internet was born in October 1969, when researchers at the University of California, Los Angeles, communicated with a computer at the Stanford Research Institute over a telephone line. This was the beginning of ARPANET, the original packet-switched network. The number of networks multiplied rapidly. In the USA, the Department of Energy established MFENet and HEPNet for its physics researchers; NASA space physicists followed with SPAN; and money from the National Science Foundation (NSF) established CSNET for the (academic and industrial) computer science community. Most importantly, the US NSFNET and UK JANET programmes in the mid-1980s were established to serve the entire higher education community. In 1990, the ARPANET was retired and transferred to the NSFNET. The NSFNET soon connected to the CSNET, and then to the EUNET, which connected research facilities in Europe.

The growth of the Internet—which in 1990 was comprised of 300,000 host computers—was fuelled by the advent of the World Wide Web (WWW). (For a full history of the WWW, see http://www.w3.org/.) The WWW is a network of sites that can be searched and retrieved by a special protocol known as a hypertext transfer protocol (HTTP). The protocol simplifies the writing of addresses and automatically searches the Internet for an address and calls up the requested document

for viewing. The idea behind the WWW is expressed most clearly by Tim Berners-Lee who invented the WWW in CERN, Geneva:

> The Internet . . . is a network of networks. Basically it is made from computers and cables. . . . The Web is an abstract (imaginary) space of information. On the Net, you find computers—on the Web, you find documents, sounds, videos, . . . information. On the Net, the connections are cables between computers; on the Web, connections are hypertext links. The Web exists because of programs which communicate between computers on the Net. The Web could not be without the Net. The Web made the Net useful because people are really interested in information . . . and don't really want to have to know about computers and cables.

In 1995 the NSFNET backbone—the portion of the NSFNET used for large-volume, long-distance transmission—reverted to a research network. US Internet traffic requiring backbone transport was routed through the networks of the private Internet providers. This privatization translated the Internet into a commercial enterprise spanning the globe, with 6.6m hosts spread over 61,000 networks with 23,500 web sites. Since 1995, the growth has continued at a remarkable rate: the number of hosts has grown at an annual average of over 60 per cent, and there were over 22m web sites by the end of 2000.

III. CURRENT ORGANIZATION OF THE INTERNET

(i) Technical Description

A standard telephone call is executed by establishing a link between two individuals that is exclusive to the call and lasts for its duration. This system is known as circuit switching. At the beginning of the 1960s, the idea of packet switching was developed. This involves breaking a call, or other message, into individual pieces of information, or packets. Each packet is then sent independently to the destination through the network, and the entire message is reassembled when all the packets arrive. No connection is established: there is no end-to-end circuit as there is for a standard telephone call. (For more discussion, see Mackie-Mason and Varian, 1997.)

This has two major implications. First, it means that network resources can be shared more effectively. A typical telephone conversation has many long pauses punctuated by short bursts of data (speech). Because the circuit used for the call is dedicated, there is no way to use the idle capacity during the pauses. When communication occurs with packets, however, many different calls can be placed over the same network, with the packets from one call being transmitted in the pauses of other calls. This process is known as multiplexing. Second, it makes communication more secure by removing the dependence of any call on any particular communication link. If a link in a network stops working for some reason, then the packets are simply re-directed through another route. (An important part of the Internet, which is not covered here, is the system of computers, or routers, that deals with the directing of packets.)

The Internet protocol (IP) provides for addressing and forwarding of individual packets. Every computer attached to the Internet has a unique IP address that enables other computers to find it. An IP address is made up of four numbers between 0 and 255, commonly shown separated by periods. For example, the IP address of the Oxford University Press is 130.88.203.71. At the top of each packet, in the 'header', is stored the IP addresses of the computer sending the packet (the source) and the computer receiving the packet (the destination). In the words of Vint Cerf, one of the pioneers of the Internet:

> a packet is a bit like a postcard with a simple address on it. If you put the right address on a packet, and gave it to any computer which is connected as part of the Net, each computer would figure out which cable to send it down next so that it would get to its destination.

Consider what it takes to send a single e-mail. The average e-mail message is 18,500 bytes (see http://www.sims.berkeley.edu/research/projects/how-much-info/internet.html; a byte is a measure of the quantity of information; the phrase 'what is a byte?' is 16 bytes long). A typical packet length is about one thousand bytes, so that the average e-mail is broken into around 20 packets. Each packet must make its way from the source to the destination, and be reassembled in the correct order once it has arrived, in order for the e-mail to be sent successfully. To achieve this, the header of each packet contains not just the IP addresses of the source and destination,

but also the size of the packet, the total number of packets in the complete message, and the number of the packet in the whole sequence of packets making up the e-mail. These data all provide the information needed to transmit the e-mail as a sequence of packets and reassemble the sequence at the destination.

(ii) Market Structure

The previous section gave a largely technological answer to the organization of the Internet. This section gives an alternative view, describing the companies and organizations that own and control the key components of the Internet.

For most residential users, access to the Internet uses a standard telephone line, using a modem to convert the computer's digital information to the analog waves that telephone lines transmit. Others may use a cable television network requiring a specific modem. Business users are connected via a local area network (such as a campus network at universities), comprised of infrastructure owned by the institution or leased from telecommunications firms. A major difference between residential and business users is that the former's connection is intermittent, while the latter is always connected to the Internet.

For most residential users, an Internet service provider (ISP) provides access to the Internet whenever the user calls the telephone number used by the ISP for dial-up access. For other users the ISP provides a direct connection from the local area network to the Internet. The ISP market is very diverse. At the end of 1999, there were over 4,000 ISPs in operation in Western Europe; there were around 7,700 ISPs in the USA in 2000 (see www.isp-planet.com/research/isps_western_europe2a.html and http://news.cnet.com/news/0-1004-200-2889725.html). There are over 700 ISPs serving the UK, and six serving Iceland (see http://new-website.openmarket.com/intindex/00-06.htm).

ISPs vary hugely in size and type. One classification divides the market into three types. First there are local or regional ISPs: small, in terms of the number of subscribers (typically only a few hundred), the range of services offered (mostly access to the Internet plus basic e-mail and Web services for residential users), and the extent of their infrastructure (typically, these ISPs do not own the facilities—computers and switches—that Internet access requires, and may have limited capacity for the number of simultaneous users). They are often called Internet access providers (IAPs) to emphasize that they provide access but little service. Second, there are national-scale ISPs, with between a few thousand and a few hundred thousand subscribers, a larger range of services (to both residential and business users), and often their own access facilities. Finally, there are international ISPs, with millions of subscribers, a wide range of services for many different types of users, and often, but not necessarily, extensive proprietary infrastructure. This last group of ISPs contains the names that are likely to be most familiar: some own their own infrastructure, others (such as AOL) do not.

While there is a large number of ISPs, the market is highly concentrated. The top six providers account for over 73 per cent of the US market by subscribers. Only 20 per cent of the ISPs in the USA operate nationally, but they generate 80 per cent of total revenues (see http://www.isp-planet.com/research/census_q12k.html and http://cyber-atlas.internet.com/big_picture/hardware/article/0,1323,5921_471621,00.html). The smaller ISPs rely heavily on the larger networks to ensure connectivity to the Internet. The international ISPs are the motorways of the Internet and are often referred to as Internet backbone providers (IBPs) to emphasize their role. There is no hard-and-fast rule as to who qualifies as an IBP, and so the number of IBPs is anywhere between five and 50.

IV. REGULATION AND COMPETITION ISSUES

Here we focus first on regulatory or competition issues associated with *conveyance* on the Internet. Except for those in large organizations, such as multi-national firms or higher education institutions, users rely for their Internet access on a copper or co-axial cable which delivers additional conventional regulated services, such as voice telephony or cable television. The regulatory framework for these services thus abuts on Internet access. Even where the conveyance of Internet traffic breaks

free of the public switched telephone network (PSTN), the growing tendency to carry voice traffic on data networks, though use of the VoIP (Voice over Internet protocol) brings the two together again.

(i) Dial-up Access

In most countries, both the retail price of local calls supplied by the dominant operator and the wholesale price of call origination and call termination are subject to price regulation. This is considered necessary because of the dominant position held by the historic operator in the provision of local telecommunications service (including call origination) and the supposed bottleneck property of call termination. As a result, when PSTN subscribers seek access to their ISP, they do so within a framework of telecommunications regulation.

The simplest form that this access can take is when a customer on the dominant operator's network gains access to an ISP via that network and all call revenues are retained by the network operator. Alternatively, the subscriber's network operator passes the call to the ISP's operator, which provides a termination service. As before, the retail price is subject to regulation, but the further question arises of how that regulated call revenue is divided among the parties (the originating operator, the terminating operator, and the ISP). The key issue is whether the originating operator (as would be the case with a normal call to another subscriber) keeps the revenue and pays a call termination charge to the operator to which the call is passed, or whether the roles are reversed so that the revenue accrues to the terminating operator, which then pays an origination charge to the subscriber's operator.

The second option may come into play in Internet access, because access to ISPs is often achieved via a range of specially tariffed services primarily used for telemarketing, including, in the UK, free phone services (0800 numbers), local call fee access (0345 numbers), national call fee access, and premium rate services.

Much of the debate in the UK about Internet access has arisen as a result of the method for dividing revenue associated with BT's Number Translation Services adopted by Oftel in 1995. The formula is as follows (Oftel, 1999):

$$\begin{aligned} &\text{the originating operator retains: } P - D + C \\ &\text{the terminating operator receives: } D - C \end{aligned} \qquad (1)$$

where P is the actual retail price charged by the originating operator to the customer; C is the pence-per-minute charge for conveyance over a single tandem segment of BT's network, including an uplift to allow for retail costs incurred by the originating operator in handling the calls; D is the deemed retail price for the call. In the case of free telephone services this is 0. In the case of local and national call fee services, it is the retail price adjusted for discounts.

Under this formula, the terminating operator keeps any surplus over marginal cost (where that marginal cost includes the origination charge). However, the process may not end there.

(ii) Internet Access and the Free ISP

In late 1998, Freeserve, a fully owned subsidiary of Dixons, a large electrical retailing firm, made available for the first time in the UK a free ISP service. Customers could simply take a free disk from Dixons stores or elsewhere, and load it into their personal computer (PC). This gave them access to Freeserve via a local call charge number, with no charge for the basic ISP service for which its competitors charged a monthly fee.

Freeserve quickly acquired over a million customers, and became the largest ISP in the UK. Others, followed its lead. The free ISP business model, as originally developed in the UK and subsequently adopted in many other counties, is essentially a regulatory artefact: because the share of the retail price accruing to the terminating operator excludes that terminating operator's costs, there is an opportunity for the ISP to appropriate the rent, by shopping around for a terminating operator.

However, competition in the ISP sector forces them to re-cycle it back to customers in the form of a zero ISP charge. From a consumer's point of view, this is obviously preferable to an alternative in which the rent remained either with the telecommunications operator or with an ISP. However, other arrangements which might benefit consumers are possible. For example, the terminating operator could reduce both its termination rate and the retail call price paid by the subscriber.

This would re-cycle the rents associated with access to the ISP at the standard call rate back to the consumer directly, rather than indirectly via the ISP. Or the regulator could simply reduce, or alter the structure of, the retail price for Internet access, differentiating it from the retail price of voice calls. This has been done in a number of member states of the European Union, and, implicitly, in the United States, where the Federal Communications Commission (FCC) has exempted ISPs, viewed as a special class of enhanced service providers, from paying access charges to local exchange carriers for use of their network. Or optional calling plans can be developed with quantity discounts especially for calls in off-peak periods, which approximate to the North American model of unmetered local service, the presence of which is partly held responsible for high US Internet penetration rates.

(iii) Regulating Flat-rate Internet Access Call Origination (FRIACO)

Not surprisingly, European ISPs and the telecommunication operators they choose to terminate their calls have been agitating for the availability of a flat-rate Internet access call origination product, on the basis of which they can offer flat-rate retail tariffs. Under the terms of the EU Interconnection Directive, the incumbent telecommunications operator can be called upon to supply such a product. In cases where that operator is offering its own flat-rate retail service, such as BT's Surftime, the obligation to provide a wholesale equivalent is even stronger.

However, the process of designing and pricing such a product has proved difficult and protracted. In some countries, the incumbent has argued that flat rates carry dangers for the network as a whole, which might be unable to cope with the resulting traffic and might even suffer catastrophic failure. If that obstacle can be overcome, there are still quite serious technical difficulties in predicting the pattern of usage by flat-rate subscribers and hence the cost of such access. Indeed, since subscribers will self-select into per-minute and flat-rate groups on the basis of the retail prices, which are themselves strongly influenced by the wholesale price, the task of prediction is made more complicated by circularity.

In the UK, Oftel developed a procedure for costing flat-rate interconnect access. (Oftel, 2000). Because the costs of the telecommunications network, especially the capital costs, are driven primarily by capacity rather than by minutes of usage, there is no difficulty in principle in establishing the cost of a particular circuit connecting a subscriber to the local exchange. However, an operator buying call origination from BT would not necessarily have to purchase a circuit for each customer, except in the unlikely event that each customer were using the circuit 24 hours a day. Normally, the operator would aggregate customers' needs, so that each circuit satisfied a number of them. There would, however, be a risk in measuring the demands placed on BT's network simply by computing the maximum volume of traffic handed over by BT to the terminating operator's interconnection port, because the total amount of capacity which BT would have to install would also depend upon how demand was distributed at other points in the network; a given capacity at the point of interconnection might require different levels of investment in other parts of the network, if the busy hours in those other parts of the network were non-coincident. As there is no reason to suppose that the distribution by time of day of Internet traffic would exactly match that of existing voice traffic, it is hard to forecast these patterns of additional demand, and hence the costs of network expansion. These uncertainties forced Oftel to adopt an interim approach to pricing, even though the logic of costing FRIACO is well established.

(iv) Broadband Access

The discussion so far has focused on narrow-band Internet access—to which most users currently have access. Increasingly, however, customers are seeking a much faster service made available through broadband access, which can be provided both by cable modems and by an adaptation to the standard telephony copper wire known as ADSL (asymmetric digital subscriber line). These technologies are subject to different regulatory arrangements.

The operator of a telecommunications network can offer ADSL by placing appropriate equipment in the subscriber's home and in the exchange. This enables the subscriber to receive high bit-rate transmissions—for example of video programmes from a

server or data from websites. The bit-rate depends upon the subscriber's distance from the exchange and other technical factors, but it is estimated that in the UK a service offering video of a quality equal to terrestrially transmitted television could be made available to about half of telephone subscribers.

The question arose as to whether, as an extension of the standard mandatory access which other telecommunications operators have on a per-minute basis to the incumbent's network, competing operators were entitled to lease the incumbent's local loop. In the United States, this was mandated under the 1996 Telecommunications Act, as part of a general requirement for unbundling. In the European Union, the legal position under the Interconnection and other Directives was unclear, until 2000 the European Parliament's Council of Ministers adopted a mandatory regulation requiring unbundled access to the local loop. National regulatory agencies have approved, or are currently in the course of approving, prices for such access. The process of unbundling the loop in the United Kingdom was protracted, delayed, and controversial. In other member states, and in the United States, unbundled loops have been available for several years, although the take-up rate is a fraction of 1 per cent of lines.

Debate has also raged about whether cable companies should be obliged to offer access to their networks. In both Europe and the United States, cable operators are under no such general obligation, and hence can limit access to their subscribers by competing ISPs. In the United States, the regulatory authorities have been able to unlock access to the largest cable system as part of the approval process for the Time Warner/AOL merger. In Holland, the Dutch parliament has passed legislation requiring the regulator to ensure mandatory access within a specified period. In the UK, Oftel set out its policy on open access in April 2001, essentially adopting a competition policy approach. This involves establishing whether the operator of the network in question possesses power in the relevant market, being satisfied that the expected benefits exceed the costs and that open access is the most effective and proportionate measure available. (Oftel, 2001). The decision would hinge critically on market definition, in particular the questions of whether narrow-band and broadband access fell in the same market, and whether broadband access technology, such as ADSL and wireless techniques, exercised a competitive constraint on cable networks. (This issue is discussed in the US context by Speta (2000).)

At present, however, competition issues are secondary in the UK public policy debate to the slow roll-out of broadband. In 2001, the UK was one of the poorest performers among OECD countries in broadband penetration, and retail prices were increasing. This sat uncomfortably with the government's target for 'broadband Britain' and is a matter of major concern

(v) Wireless Access

The development of wireless application protocol (WAP) technologies has already permitted Internet access using second-generation mobile telephony. This is particularly popular in Japan, under the name of i-mode. Fast Internet access will be available under third-generation wireless technology. Licences to provide such services have been allocated in Europe and elsewhere since 1999, earning eye-catchingly large amounts for some governments—up to 5–600 euros per head of population. In the UK a subsequent auction for licences to provide fixed wireless internet access drew disappointingly few bids, and many of the lots failed to reach their reserve price. Internet access can also be made available by satellite, typically accompanied by a low-capacity telephone-based return-path.

There are, in addition, prospects for using other frequencies for wireless local area networks, which can provide inexpensive Internet access. Wireless technologies thus can both add additional features, such as mobile access, and may also have cost advantages for access in sparsely populated regions.

(vi) Competition Law Approaches to Backbone Networks

At the other end of the scale from residential internet access lie the high-capacity networks which most ISPs use to achieve global interconnection for their customers. The supply of high-level connectivity has been at issue in two merger investigations carried out by competition authorities in the United States and Europe. The first led to the merger, with

conditions, of MCI and WorldCom in 1998. The second led to the rejection by the European Commission of a proposed merger between MCI/ WorldCom and Sprint in 2000. That decision was subsequently appealed to the European Court of the First Instance.

In the case of both mergers, Internet-related questions concerned market definition, and whether the combined entity created by the merger would be dominant. On the former question, the parties to the proposed merger argued that an ISP could purchase connectivity by a variety of means, including the purchase of transit from so-called backbone networks, the conclusion of peering (or barter) agreements with national or continental equivalents, or the purchase of transit from smaller ISPs.[2] On this argument there would be a continuous chain of substitution between the more limited connectivity offered by smaller ISPs and the global connectivity offered by the largest operator.

The Commission, however, adopted a narrower definition which distinguished top-level providers, defined by the characteristic that they held a set of peering agreements which equipped them with a very high level of settlement free connectivity across the Internet. On this basis, the combination of MCI and WorldCom in the first proposal, and of WorldCom and Sprint in the second, would have relatively high market shares. The Commission conditioned its approval of the MCI/WorldCom merger on divestment of MCI's Internet assets.

On the issue of dominance, the parties argued that barriers to entry in the market as defined by the Commission were low and that ISPs were able to switch to a competing supplier of high-level connectivity without difficulty. ISPs could also respond to higher transit prices by adopting technologies such as the replication (mirroring) and local caching of sites which permitted the substitution of regional for global connectivity.

The Commission, on the other hand, formed the conclusion that the merger would generate a tipping effect. Under this process, ISPs would gravitate toward the largest network, because it would offer a better quality of service. This would partly be due to the fact that service provided by the largest network would involve a smaller number of 'hops' or transitions between networks, which might cause delays. In addition, customers might rationally conjecture that the largest network would deliberately degrade its facilities for interconnection with its competitors, thus ensuring that their customers enjoyed a lower quality of service and drawing them on to its own network. This second argument was developed at the theoretical level by Crémer *et al*. (2000), who advised GTE on the Internet aspects of the two mergers.

(vii) Content

The purpose of the Internet is to provide access to content, either for point-to-point transmission, for example through e-mail, or for point-to-multi-point purposes, such as access to a web site. The Internet thus represents a platform for content distribution, competing with cable, satellite, and terrestrial broadcasting platforms.

So far, the Internet platform has made relatively little headway in competing with other platforms for paid content. Internet distribution rights for premium sporting events do not yet command high prices. The principal areas in which Internet-delivered content is charged are pornography, which regulation has excluded from other platforms, and games. In consequence, most Internet-delivered content is remunerated through e-commerce revenue, or by the sale of advertising. Sale of the latter is, however, made difficult by the absence of reliable techniques for monitoring the number of visitors to any site.

In relation to other platforms, competition and regulatory issues have arisen relating to vertical integration between content and delivery. Given the high level of interconnectivity which is the hallmark of the Internet, the limited barriers to putting content on the net and the relative insignificance of the Internet as a delivery platform for premium video content, these issues have not yet arisen. In both France and the UK, however, regulators have stepped in to prevent a mobile delivery operator from excluding its residential subscribers from access to portals other than those associated with the network itself.

[2] The differences between peering and transit are discussed below.

V. PROBLEMS AND PUZZLES

The previous section discussed some of the major competition issues concerning the Internet. In this section, we review the economic analysis underlying three further issues of major interest.

(i) Congestion Pricing

By any measure, the growth of the Internet has been phenomenal: the number of hosts, the number of users, and the amount of traffic have been doubling approximately every year since 1988. The price of this success has been increasing congestion. Surfing the Web is notoriously slow during peak hours; by some estimates, 30 per cent of Internet traffic is re-transmissions of dropped packets. (It is surprisingly difficult to obtain hard evidence of Internet congestion. See Paxson (1997) for an authoritative study of the area. Many university links to the public Internet are heavily loaded, which may be why academics think congestion is a problem. It may be, however, that the general problem is not congestion, but non-responding servers; see Huitema (1997).)

It is widely recognized that pricing of Internet resources is required to control congestion. Yet, as noted above, the trend is towards the flat-rate pricing and low variable prices for Internet usage which have been one of the main drivers of congestion on the Internet. The underlying economic problem is an old one, known as 'the tragedy of the commons', a termed coined by Hardin (1968) in his seminal article on the optimal use of grazing land. Users of any common and freely accessed resource have a tendency to over-exploit: each will use the resource until the private (marginal) cost of doing so equals the private (marginal) benefit, ignoring the social consequences of their actions.

Many pricing schemes have been proposed to combat the problem of rising congestion. The best-known Internet pricing scheme is the so-called 'smart market', proposed by Mackie-Mason and Varian (1997). The proposal involves a zero usage price when the network is uncongested. When the network is congested, however, packets would be prioritized based on the value that the user puts on getting the packet through quickly. Each user assigns a bid to all of his or her packets, corresponding to his or her willingness-to-pay for immediate service. At congested parts of the network, packets are prioritized based on bids. The key to the scheme is that users are not charged the price they bid, but rather pay the bid of the highest priority packet that is not admitted to the network. This scheme gives users incentives to reveal their true willingness-to-pay for priority; and it generates the socially optimal level of revenues for network expansion. The smart market is therefore an example of a second-price, or Vickrey (1961) auction.

There are several criticisms of this scheme. The first is that it fails to take into account the fact that users are interested not only in instantaneous resource allocation, but allocation over time (e.g. the duration of an Internet telephone call); for recent work on this question, see Crémer and Hariton (1999). Second, the smart market is generally viewed as being too complex to implement. The requirement that a bid is attached to every packet imposes large burdens on both users and already-congested resources (especially routers). (Recall that the average e-mail generates about 20 packets; this paper is around 200 packets-worth; a 5-minute telephone conversation generates around 1,500 packets.)

A simpler scheme is Andrew Odlyzko's 'Paris Métro Pricing' (PMP) proposal (see Odlyzko, 1997), based on the system that was used some years ago on the Paris Métro. Users of the Métro were offered a choice of travelling in first- or second-class carriages. The only difference between the two carriages was the price charged: both carriages had the same number and quality of seats, and (obviously) both reached the destination at the same time. The first-class carriage was, however, more expensive, and consequently (on average) had fewer passengers in it. Passengers sorted themselves appropriately. There are other examples (see Chander and Leruth, 1989).

Odlyzko applies the same scheme to packet-based networks, such as the Internet. His idea is to partition a network into separate logical networks, with different usage charges applied on each sub-network. As with most PMP schemes, there is no guarantee of service quality, but subscribers to the more expensive network expect lower average use rates, lower average congestion, and hence faster delivery. Users sort themselves according to their preferences for congestion and the prices charged

on the sub-networks. In fact, this sort of thing happens now. Dial-up Internet users can choose between cheap services that are slow or difficult to access, and more expensive services that offer faster rates and access.

The attraction of the PMP scheme is its simplicity—it involves only a small number of service classes and little computation is required to assign traffic to the right class. This should make implementation of the scheme easier, quicker, and less expensive than the more complex smart market proposal. Gibbens *et al*. (1998) show, however, that the PMP scheme may not survive in non-cooperative equilibrium. Their analytical findings are similar to the numerical results of Wilson (1989), who shows the same for priority supply classes of electricity. The intuition for the result is that any benefits to networks from offering multiple service classes (extraction of surplus through price discrimination) are outweighed by the increased competition that this causes (see also the recent literature on price discrimination in oligopoly: Stole (1995), Rochet and Stole (1999), Mason (2000), and Armstrong and Vickers (2001)). The main conclusion that emerges from this research is the importance of placing any pricing proposal within an economic model of network competition.

(ii) Entry to the ISP Market

A remarkable feature of the ISP market is the scale of entry. In the UK, there were around 350 ISPs in 1999; this number doubled during 2000. The 700 ISPs share between them around 9m ISP subscribers: an average of around 10,000 subscribers per ISP. Of course, few ISPs have as many subscribers as this, since a small number of ISPs account for most of the market (the 10 largest have nearly 90 per cent of subscribers). The large number of providers is surprising, given the economies of scale that arise naturally with Internet networks. A key characteristic of Internet design is traffic aggregation: traffic from many sources is combined and carried over shared lines. This sharing exploits the fact that typically traffic is 'bursty'. Someone using the Web spends a little time downloading pages and a lot of time reading them. This behaviour means that each individual generates short episodes of heavy traffic flow interspersed with long pauses.

If an ISP installed capacity to deal with the peak traffic rate of each user, it would find average usage of its facilities to be very low. Instead, ISPs install lower capacity, but combine many users whose peak traffic flows are statistically independent. In this way, the ISP can meet the peak demands of most of the users most of the time, while spending less on capacity. The greater the number of users the ISP can attract, the more it can use statistical aggregation to drive its unit cost down. Put in more standard economic terms: the nature of Internet traffic leads to increasing returns to scale. Clark and Lehr (1999) estimate that the minimum efficient scale (MES) for an ISP, arising purely from traffic aggregation, is of the order of 5,000 to 50,000 subscribers. For the UK, 690 ISPs share 1m subscribers, giving them an average size of around 1,500 subscribers, well below the lowest MES estimated by Clark and Lehr.

Two factors explain why the average ISP is well below the MES. First, in many countries, most Internet access uses telephone lines. At the beginning of 2000, 47m of the 50m online accounts in the USA used dial-up over a standard telephone line for access to the Internet; see http://www.isp-planet.com/research/census_q12k.html. The combination of a standard modem over a standard telephone line limits the rate at which users can send and receive traffic, typically to 56,000 bits per second (where 8 bits make a byte). To put this in context, recall that an average e-mail is 18,500 bytes, or 148,000 bits; this takes 2.5–3 seconds to transmit. A typical music file comes to 14m bytes (or 14 megabytes) and would take around half an hour to transmit. In contrast, an office machine connected to a local area network typically can transmit and receive at 10m bits per second, and would send a music file in seconds.

This technological cap on traffic rates limits the economies of scale that ISPs can enjoy from traffic aggregation. Since the peak rates of dial-up users are so low, their traffic flows are much smoother; consequently, the pauses are much shorter and so fewer sources can share the same line. The MES estimate of over 5,000 subscribers assumes that the ratio of peak to average rates is 100. For Web browsing, the average rate is roughly 10,000 bits per second; so the ratio of peak to average rates with the

modem/telephone line combination is around 5. This decreases the lowest MES to approximately 1,000 subscribers: close to the average size of the ISPs outside the UK's top ten.

These calculations are rough-and-ready, but they indicate the cost considerations that are relevant for analysing entry to the ISP market. As new broadband access technologies become widespread, the economies-of-scale calculation will shift and it seems likely that pressures for consolidation will mount. A detailed model of the technology and economics of local access and its effect on the ISP market would be very valuable and should be a priority for future applied research in this area.

The second factor behind the large number of ISPs is the structure of regulation in the telecommunications sector, and particularly termination charges. In the early days of Internet service provision, ISPs had to rely on subscription fees and advertising revenues to generate income. In Europe, the liberalization of the majority of telecoms markets in 1998 opened up revenue sharing as a new source of income for ISPs, as described in section IV(iii) above.

It might be expected that this new source of income would encourage entry by ISPs. The statistics certainly bear out that conclusion: for example, the number of ISPs in the UK doubled during 2000. The issue is complicated, however, by several factors. The first is the effect of the revenue-sharing arrangement on the nature of competition. The number of subscribers of an ISP becomes all-important for its bargaining power in agreeing the split of revenues with its terminating carrier. The more traffic that the ISP generates, the greater the revenue share it can command. This situation is similar, but not identical, to competition with positive network externalities. In the former, inverse demand (that is, unit price received by the ISP) is determined by the bargaining process between the ISP and the terminating carrier. In the latter, a consumer's willingness-to-pay for a good rises with the total consumption of the good; see, for example, Katz and Shapiro (1985) and Economides and Himmelberg (1995).

An implication of this observation is that, in the long run, the structure of the ISP market may become more concentrated. A major lesson from the network externalities literature is that demand-side effects provide large economic incentives for a small number of technologies, products, or firms to dominate. This is true because, given a choice between two competing alternatives with network externalities, users will choose the alternative with the largest number of users. Take the example of competing but non-interconnecting telephone systems. The system that provides links to only 25 per cent of users is less valuable than the system that connects to 75 per cent of users. People are therefore more willing to join the larger system, with the consequence that one system dominates. The network externalities, therefore, produce a winner-take-all outcome; expectations are self-reinforcing, in that the system that is expected to be larger is dominant (Shapiro and Varian, 1999). The similarity between standard network externalities and the bargaining-induced network effect suggests that the revenue-sharing arrangement may do the same. However, changes in telecoms regulation may alter the terms of revenue sharing.

(iii) Interconnection

The Internet is all about connectivity: any two computers anywhere in the world can, in principle, communicate with each other. This possibility is supported by thousands of interconnection agreements between the many separately owned communication networks that comprise the Internet. This section examines aspects of these interconnection agreements, and in particular the great peering debate.

The gulf between large and small networks has widened progressively with the commercialization of the Internet. By November 1997, it was estimated that the USA's four largest networks (UUNet, MCI, BBN, and Sprint) accounted for between 85 and 95 per cent of total backbone (i.e. core) Internet traffic, with the remaining volume carried by upwards of 40 other, small, networks; see OECD (1998). The growing asymmetry led to concerns that larger networks might discriminate against smaller rivals.

There has been very little economic analysis of interconnection in the Internet, and so there is little idea whether such concerns are warranted. There is a well-established literature considering the gen-

eral issue of compatibility; see, for example, Katz and Shapiro (1985) and Farrell and Saloner (1992) for network analyses; Matutes and Regibeau (1988), Economides (1989), and Einhorn (1992) examine compatibility without network effects. There are a few papers that look explicitly at interconnection agreements; see, for example, Baake and Wichmann (1999), Foros and Hansen (1999), Foros *et al*. (2000), and Laffont *et al*. (2000). Neither set of papers captures all of the details that are relevant and necessary for analysing the developments in Internet interconnection.

The issue currently being decided between ISPs is not whether to interconnect or not, or even whether to charge for interconnection or not. While peering typically does not involve payments between the peers, this is not the only important aspect of the arrangement. Equally important are the routes that the ISPs advertise to each other. An ISP has service contracts with customers (such as web sites) which are not ISPs, as well as with other ISPs. Traffic that is passed to other ISPs may be destined for either that ISP's customers, or for other ISPs (and their customers). A further distinction between transit and peering arrangements is based on which routes are made available in the arrangement. Two ISPs that form a transit arrangement give access to all routes, agreeing to accept traffic from each other that is destined for other ISPs as well as direct customers. Peers do not engage in transit and accept traffic from each other only if it is bound for their (non-ISP) customers.

Two factors are important. First, an ISP that is expected to be larger is likely to have greater bargaining power when negotiating with other ISPs over interconnection. At the most extreme, if the larger ISP were an upstream monopoly supplier of connection to a smaller ISP, it could make a take-it-or-leave-it offer. ISPs of similar size will have more equal bargaining positions. Second, as in the Katz and Shapiro (1985) model, if there is symmetry in interconnection agreements (that is, all ISPs form either transit or peering arrangements with all other ISPs), then the equilibrium outcome must be symmetric. If there are asymmetric interconnection arrangements and larger ISPs have market power, then the equilibrium outcome can be asymmetric, with the larger ISPs in the asymmetric equilibrium earning greater profits than would be the case in symmetric equilibrium. These two factors combined mean that larger ISPs can gain by peering with each other and offering high-priced transit to smaller ISPs. Details of a simple extension to the Katz and Shapiro (1985) model that illustrates this argument are available on request from Mason; see also Little and Wright (2000), Milgrom *et al*. (2000), and Kende (2000) for recent analyses.

Interconnection between Internet networks will continue to trouble regulators, particularly if consolidation within the industry continues. So far, most policy positions are based on a suspicion that large networks have market power; and they exercise this market power through interconnection agreements with smaller networks. There is little analysis to support this suspicion.

VI. CONCLUSIONS

In this paper, we have provided a description of the Internet, its structure, and background, an examination of the policy issues that are of current concern for regulators, and a discussion of the economic models that need to be developed to address these policy issues. We have come to the following conclusions.

- The Internet is a global network of networks.
- Most access to the Internet occurs over public switched telephone networks (PSTN).
- Regulation involving pricing of and access to the local loop therefore has a major impact on the Internet, and particularly the Internet service provider (ISP) market.
- Two factors are particularly important: (i) the wedge between retail and interconnection charges; and (ii) the capacity of access technologies.
- In Europe, rents, created because regulated retail prices exceed termination charges, in the short run have encouraged entry by ISPs; in the long run, the ISP market is likely to become more concentrated.

- Dial-up access over the PSTN currently limits economies of scale; new broadband access will allow economies of scale to be exploited, which will make the ISP market more concentrated.
- Congestion is a growing problem on the Internet; current pricing structures are unlikely to be optimal.
- Pricing schemes have to be assessed within an economic model of the ISP market.
- There has been little detailed economic analysis of interconnection agreements; this should be a priority for future theoretical and applied research, and should include the question of market power in interconnection negotiations.

REFERENCES

Armstrong, M., and Vickers, J. (2001), 'Competitive Price Discrimination', Nuffield College, Oxford, mimeo.

Baake, P., and Wichmann, T. (1999), 'On the Economics of Internet Peering', *Netnomics*, **1**(1).

Chander, P., and Leruth, L. (1989), 'The Optimal Product Mix for a Monopolist in the Presence of Congestion Effects: A Model and Some Results', *International Journal of Industrial Organization*, **7**(4), 437–49.

Clark, D., and Lehr, W. (1999), 'Provisioning for Bursty Internet Traffic Implications for Industry and Internet Structure', MIT, Laboratory for Computer Science, mimeo.

Crémer, J., and Hariton, C. (1999), 'The Pricing of Critical Applications on the Internet', University of Toulouse, mimeo.

— Rey, P., and Tirole, J. (2000), 'Connectivity in the Commercial Internet', *Journal of Industrial Economics*, **48**(4), 433–72.

Economides, N. (1989), 'Desirability of Compatibility in the Absence of Network Externalities', *American Economic Review*, **79**(5), 1165–81.

— Himmelberg, C. (1995), 'Critical Mass and Network Size with Application to the US Fax Market', Discussion Paper No. EC-95-11, Stern School of Business, New York University.

Einhorn, M. A. (1992), 'Mix and Match Compatibility with Vertical Product Dimensions', *Rand Journal of Economics*, **23**(4), 535–47.

Farrell, J., and Saloner, G. (1985), 'Standardization, Compatibility, and Innovation', *Rand Journal of Economics*, **16**, 70–83.

— —(1986), 'Installed Base and Compatibility Innovation, Product Preannouncements, and Predation', *American Economic Review*, **76**, 940–55.

— —(1992), 'Converters, Compatibility, and the Control of Interfaces', *Journal of Industrial Economics*, **40**(1), 9–35.

Foros, O., and Hansen, B. (1999), 'Competition and Compatibility among Internet Service Providers', Telenor, mimeo.

— Kind, H. J. and Sorgard, L. (2000), 'Access Pricing, Quality Degradation and Foreclosure in the Internet', University of California, Santa Barbara, Working Paper 7-00.

Gibbens, R., Mason, R. A., and Steinberg, R. (1998), 'Multiproduct Competition between Congestible Networks', University of Southampton, Discussion Paper in Economics and Econometrics No. 9816.

Hardin, G. (1968), 'The Tragedy of the Commons', *Science*, 162.

Huitema, C. (1997), 'The Required Steps towards High Quality Internet Services', unpublished Bellcore Report.

Katz, M. L., and Shapiro, C. (1985), 'Network Externalities, Competition, and Compatibility', *American Economic Review*, **75**(3), 424–40.

Kende, M. (2000), 'The Digital Handshake: Collecting Internet Backbones', OPP Working Paper No. 32, Federal Communications Commission.

Laffont, J. J., Marcus, S., Rey, P., and Tirole, J. (2000), 'Internet Interconnection and the Off-Net-Cost Pricing Principle'', University of Toulouse, mimeo.

Little, I., and Wright, J. (2000), 'Peering and Settlement in the Internet: An Economic Analysis', *Journal of Regulatory Economics*, **18**(2), 151–73.

Mackie-Mason, J. K., and Varian, H. (1997), 'Economic FAQs about the Internet', in L. W. McKnight and J. P. Bailey (eds), *Internet Economics*, Cambridge, MA, MIT Press.

Mason, R. A. (2000), 'Simple Competitive Internet Pricing', *European Economic Review*, **44**(4–6), 1045–56.

Matutes, C., and Regibeau, P. (1988), '"Mix and Match": Product Compatibility without Network Externalities', *Rand Journal of Economics*, **19**(2), 221–34.

Milgrom, P., Mitchell, B., and Srinagesh, P. (2000), 'Competitive Effects of Internet Peering Policies', in I. Vogelsang and B. Compaine (eds), *The Internet Upheaval*, Cambridge, MA, MIT Press, 175–97.
Odlyzko, A. (1997), 'A Modest Proposal for Preventing Internet Congestion', AT&T Labs, research mimeo.
OECD (1998), 'Internet Traffic Exchange: Developments and Policy', Discussion Paper DSTI/ICCP/TISP(98), Paris, Organization for Economic Cooperation and Development.
Oftel (1999), 'Consultation Paper on the Relationship between Retail Price and Interconnection Charges for Number Translation Services', London, Oftel.
— (2000), 'Determination Relating to a Dispute Between British Telecommunications and WorldCom Concerning the Provision of a Flat Rate Internet Access Call Origination Product ("FRIACO")', London, Oftel.
— (2001), 'Open Access: Delivering Effective Competition in Communications Markets', London, Oftel.
Paxson, V. (1997), *Measurements and Dynamics of End-to-end Internet Dynamics*, Ph.D. thesis, Computer Science Division.
Rochet, J. C., and Stole, L. (1999), 'Nonlinear Pricing with Random Participation Constraints', University of Chicago, GSB, mimeo.
Shapiro, C., and Varian, H. (1999), *Information Rules*, Harvard Business Studies.
Speta, J. B. (2000), 'Handicapping the Race for the Last Mile: A Critique of Open Access Rules for Broadband Platforms', *Yale Journal on Regulation*, **17**(1), 39–91.
Stole, L. (1995), 'Nonlinear Pricing and Oligopoly', *Journal of Economics and Management Strategy*, **4**(4), 529–62.
Vickrey, W. (1961), 'Counterspeculation, Auctions, and Competitive Sealed Tenders', *Journal of Finance*, **16**, 8–37.
Wilson, R. B. (1989), 'Efficient and Competitive Rationing', *Econometrica*, **57**(1), 1–40.