SIEPR Discussion Paper No. 06-04

Economic policy analysis and the Internet:

Coming to terms with a telecommunications anomaly

By
Paul A. David
Stanford University &
Oxford Internet Institute
October 2006

Stanford Institute for Economic Policy Research
Stanford University
Stanford, CA 94305
(650) 725-1874

Economic policy analysis and the Internet:
Coming to terms with a telecommunications anomaly

By

Paul A. David

Stanford University & Oxford Internet Institute

Previous draft: 18 May 2006
This version: 25 October 2006

Forthcoming in the
Oxford Handbook on Information and Communication Technologies,
edited by C. Ciborra, R. E. Mansell, D. Quah and R. Silverstone,
Oxford: Oxford University Press, Fall 2006.

ABSTRACT

The significant set of public policy issues for economic analysis that arise from the tensions between the 'special benefits' of the Internet as a platform for innovation, and the drawbacks of the "anomalous" features of the Internet viewed as simply one among the array of telecommunications systems, is the focus of discussion in this chapter. Economists concerned with industrial organization and regulation (including antitrust and merger law) initially found new scope for application of their expertise in conventional policy analyses of the Internet's interactions with other segments of the telecommunications sector (broadcast and cable television, radio and telephone), and emphasized the potential congestion problems posed by user anonymity and flat rate pricing. Policy issues of a more dynamic kind have subsequently come to the fore. These involve classic tradeoffs between greater efficiency and producer and consumer surpluses today, and a potential for more innovation in Web-based products and service in the future. Many such tradeoffs involve choices such as that between policies that would preserve the original 'end-to-end' design of the original Internet architecture, and those that would be more encouraging of market-driven deployment of new technologies that afforded ISPs with greater market power the opportunity to offer (and extract greater profits from) restricted-Web services that consumers valued highly, such as secure and private VOIP.

Contact Author: P. A. David
Department of Economics, Stanford University, CA 94304-6072
pad@stanford.edu ; paul.david@oii.ox.ac.uk

Economic policy analysis and the Internet:
Coming to terms with a telecommunications anomaly

1 Introduction: The evolving nature and scope of 'Internet economics'

Everyday life in the world's economically advanced societies has been touched and in some parts substantially altered by the advent of the Internet. Some among the developing economies also have felt the impacts of the explosive growth in global connectivity and the astounding proliferation of diverse innovations in applications software. These marvels distinguish the performance of the Internet as a communication infrastructure from that of its historical predecessors - such as the telegraph and telephone networks. It would have been truly remarkable had economists not been drawn to study the implications of this novel technology for creative expression, commercial activities and the organization of material life more generally.

Those economists who focus on industrial organization and regulation of industry (including antitrust and merger law) have found new scope for application of their expertise to issues arising from the Internet's consequences for the entire telecommunications sector. Another, more macroeconomically-oriented segment of the profession was absorbed, first, by the puzzling disjuncture between signs of rapid innovation in digital information technologies and the sluggish growth of productivity in the US economy - the so-called 'computer productivity paradox'. Then, in the late 1990s they found themselves absorbed in trying to understand what role (if any) had been played by the Internet in the sudden, and for many rather surprising productivity upsurge which has now outlived the bursting of the dotcom bubble and the post-2000 recession. In recent years, various debates concerned with the nature and implications of the digital divide - in both its domestic and its international aspects - have engaged attention in still other quarters of the discipline, especially among researchers concerned with distributive inequalities and policies intended to address these.

As scholars responding in their professional capacities to the developments that followed from the privatization and commercialization of the Internet in the mid-1990s, economists manifested an understandable inclination to 'lead from strength' - that is, to discuss matters that they had already worked out thoroughly in other contexts of application. In academic economics circles during the 1990s emphasis was placed on counteracting the hype of media and industry presentations of the Internet as a revolutionary and transformative technological advance, primarily by showing how the familiar concepts and tools of microeconomic analysis could be used to illuminate the commercial developments that were emerging in cyberspace.

This approach to the economics of the Internet drew on prior research that dealt with the special properties of information goods on the one hand, and on the extensive literature in the field of industrial organization that was concerned with competition and the regulation of telecommunications industries on the other. Key topics featured in the early 'positive' economics of the Internet literature included:

- the pricing of information goods and services sold online, and business strategy in markets characterized by network externalities;

- digital payment systems and electronic banking;
- the effects of particular instantiations of e-commerce such as financial brokerage and auctions;
- the pricing of access to the network backbone;
- regulation of Internet interconnection agreements and their implications for competitive entry and market structure.

The latter topics occupied pride of place among the Internet policy issues on which economists initially focused, in good part because the regulation of network access had emerged as a prominent subject of analysis in the era following the break-up of the Bell System and the movement to liberalization of telecommunication markets. McKnight and Bailey (1997) provide an influential collection of papers reflecting the leading edge of research on 'Internet economics' in the mid-1990s. The programmes of the annual Telecommunications Policy Research Conference during 1995-1999 reveal the same pattern of concentration on the part of the participants drawn from university faculties of economics in the United States.[1] Although the canvas of research has broadened subsequently, the foregoing remain core topics of university courses on 'the economics of the Internet'.[2]

The general thrust of this approach to the economics of the Internet, therefore, has been that of 'naturalizing' the subject matter by focusing on those generic features that were common to broader categories of economic activity and public policy, particularly those affecting the telecommunications sector. There were, however, several respects in which it was recognized that the new communications infrastructure differed from its precursors in significant respects, and moreover, that these differences impeded an immediate application of the preexisting corpus of telecommunications economics that had been formulated in the context of connection-oriented public switched telephone systems. The differentiating features manifested themselves in a variety of awkward, or anomalous performance dimensions that became apparent when the 'network of networks' was thrown open to general public and commercial traffic in the mid-1990s.

Most salient among them, at least for economists, were the difficulties of pricing the usage of bandwidth in order to reduce delays in transmission arising from congestion, and of establishing commercial ventures based on the same 'fee-for-service' business model that telephone operators were able to support. Related, but of less immediate economic interest, were the difficulties of providing users with services that permitted blocking delivery of nuisance messages (spam), offensive content, or politically disturbing material (whether of the neo-Nazi or Falun Gong variety); and protecting users from damage resulting from the malicious actions of others (for example, release of destructive software viruses, and denial of service attacks on web servers). Economic analysts could readily construe all of these as posing challenges for the design of resource allocation mechanisms that would render the Internet more efficient as a system of communications.

What was not equally appreciated, however, was that each of those problematic aspects of the Internet was rooted in the technical specifications that were responsible for the performance capabilities of the Internet, which its users perceived to be its uniquely beneficial properties. The openness and transparency of the softly integrated 'network of networks' were attributes that derived from the distinctive 'end-to-end' design of the architecture and transmission control mechanisms of the

new, connection-less communications system. These features enable the Internet to tolerate extreme diversity in the technical specifications of its constituent networks and platforms. That, in turn, made joining the system cheap, and highly attractive to new network operators, Internet service providers (ISPs) and users. In addition to facilitating the rapid extension of the system, the transparency of the end-to-end architecture - which placed the intelligent components at the ends rather in the core of the network - afforded a particularly accommodating platform for developers of applications innovations (David 2001a). Software could be designed to run on the computers situated at the network's edges, taking data input and generating data output that traversed the intervening channels without having to pay attention to the specifics of the computer hardware and software that executed the message routing functions of the system.

The existence of a significant set of public policy issues for economics to address, arising from the tension between the 'anomalous drawbacks' and the 'special benefits' of the Internet, is the focus of this chapter. Yet, for a considerable number of years it was not recognized explicitly in the emerging literature on Internet economics. One hesitates to fault a discipline that disposes its expert practitioners to talk about things they really do understand - and what the first economists to enter the field understood well was the pre-Internet world of telecommunications. The question is how far it is possible to go on the basis of understandings gained in contexts that have some similarities, but within which the subject of interest appears quite anomalous. I suggest that economists working in this area have tended - for rather too long - either to avoid focusing on the points of divergence between connection-oriented and connection-less communications systems, or to propose 'solutions' to perceived inefficiencies that would have the effect of bringing the economics of the Internet more closely into line with that of the class of telecommunications systems with which they were already familiar.

The understandable inclination of analysts to focus on features of a new the phenomenon that allow them to speak authoritatively may fail adequately to address core policy issues that are posed by radical innovations in technology - or in institutional design, for that matter. Therefore, I shall not review here the many important questions on which economists examining the Internet have been able to proceed quite usefully by recapitulating their prior concerns, and deploying familiar and well-honed tools to illuminate new developments. These have been ably surveyed elsewhere (Graham 2001). Instead, this chapter is directed to some policy issues that were avoided by initial forays into the economics of the Internet, and whose importance only lately has begun to come into clearer view.

The discussion will proceed through the following steps. In the second section it will be shown, first, that the problem of congestion which was widely perceived to be a critical economic resource allocation challenge turned out to be largely chimerical; and, second, that economists were quite blasé in proposing solutions for that and other, related problems by introducing pricing mechanisms whose implementation required radical engineering modifications to the Internet. The latter, however, would jeopardize and possibly vitiate the unique, socially valuable performance capabilities of the system. The third section takes note of the fact that the recommendations for usage-sensitive pricing that were advanced by economists on static and rather narrow 'efficiency' grounds probably posed less of an actual threat to the continuation of the end-to-end design principle than the pressures that presently emanate from the private sector. The source of the latter are business ventures seeking

the engineering modifications needed to enable them to offer users more profitable services, such as voice telephony over the Internet. The fourth section therefore addresses the questions of whether and on what grounds public policy might be mobilized to protect the architecture of the Internet - and thus preserve its beneficial properties of flexibility, extensibility, and hospitality as a platform for innovation. The final section concludes with some observations and suggestions regarding future policy-relevant directions for Internet economics.

2 Remedies for 'an unpriced resource'

Among the early contributions to the economics of the Internet perhaps the best known were those concerned with the sources of congestion, and how to deal with them (Mackie-Mason and Varian 1996; 1995a; 1995b). What economists typically brought to this discussion, perhaps all too predictably, was an abstract understanding of the phenomenon of congestion as a negative externality suffered by all users as a consequence of the lack of some effective mechanism restraining individuals' claims on the limited available capacity. Casual analogies were drawn with the phenomena of overfishing and overgrazing of common resources, and the spectre was thus raised of the Internet becoming another case of a resource whose utility was seriously degraded by congestion arising from the absence of (bandwidth) usage-sensitive pricing. The mantra that subsequently has been imparted to novitiates in the field of Internet economics carries the same message, formulated in a less normative way (McKnight and Bailey 1997: 12): 'Flat rate pricing does not provide an economic congestion control mechanism for bandwidth resource allocation'.

Most of the proposals put forward by economists to correct this deficiency have favoured usage-pricing and a useful review is provided by Gupte (2001), although their schemes have varied considerably both in the degree of their economic sophistication and their complexity. At the upper end of that scale, the 'smart market' mechanism advocated in the pioneering work of Mackie-Mason and Varian (1995a) applies the principles of a Vickery auction: users would enter bids for network access that indicated a maximum willingness to pay, and routers would recognize the bids attached to each of the data-packets; all packets with bids exceeding some cutoff value would be admitted for forwarding. Given a fixed supply of bandwidth, the cutoff value would therefore be the lowest bid that corresponded to the transmission capacity of the system, and that price would be charged to all users whose bids were accepted. Consistent with marginal cost pricing principles, when there were only bids for network access that fell below the router's cutoff value, the price would fall to zero.

As the authors of this proposal soon acknowledged (Mackie-Mason and Varian 1995b): "usage-based pricing is itself expensive - it requires an infrastructure to track usage, prepare bills, and collect revenues." A subsequent publication (Mackie-Mason and Varian 1997) took the matter further by recognizing that designing a congestion accounting and billing mechanism for a packet network is not so straightforward a proposition: Who should be charged, the sender of packets, or the receiver? Consider the situation in which a user downloads a file from a public archive: both the applications that are parties to the communication-transaction originate their own packets, but there is no way for the routers to identify the many packets forwarded from the archive as being responses to the session initiated by the small number of packets carrying the user's request for the file. If such requests

resulted in congestion, how could the behaviour of the users be modified by charging the costs to the passive party in the transaction (the archive)? To allocate the congestion costs between the parties, the public archive in this case would have to have installed a billing mechanism, permitting the subsequent reassignment of the charges to the user that had instigated the file transfer.

Just what changes would be required in the architecture and transmission control algorithms to enable the routers to do all this was not considered. But, the design of the Internet's transmission control protocols (TCP) does not allow monitoring the state of congestion everywhere in the network, and so the implementation of the suggested pricing mechanism, like that of quality of service (QOS) schemes, would require monitoring and information collection functions that are not supported and - with the continuing growth of the network - would become increasingly taxing for the simple routers to accomplish in real time. Moreover, the cost allocation and billing requirements for congestion control via QOS systems would call for the collection, transmission, and processing of *internal* traffic information as well as user bids, and the provision of discretionary network routing capabilities. To imagine all that being implemented without substantial engineering departures from the principles of an end-to-end architecture is difficult indeed (Odlyzko 1998: 26-7; CSTB 2001: 99-100), and so it seems rather remarkable that the larger implications of such changes have not been more prominent matters of concern for the proponents of such schemes.

More remarkable still is the continuing robustness of the economics literature's fixation on congestion-pricing, the pertinent facts notwithstanding. Congestion was not a major problem on the Internet during the early 1990s, when its opening to commercial traffic first directed attention to the problem posed by the impending need to introduce usage-pricing; nor has the forecast condition of chronic congestion materialized subsequently. Delays experienced on the Internet will indeed be caused by queues, which are an intrinsic part of congestion control and the sharing of capacity (CSTB 2001: 98ff). But there can be other sources of delay. Indeed, because ISPs are not required either to collect or release data on transmission delays, dropped packet rates, or other network performance variables, there continues to be much disagreement over the exact extent to which many of the service problems experienced by Internet users are properly attributable to congestion, rather than other causes. The frequently observed delays in the delivery of e-mail, for example, are thought to be almost always the result of mail server faults that result in a large proportion of the load being generated by the retransmission of packets; and the painful slowness that web surfers encounter during peak hours is ascribed to nonresponding web servers (Odlyzko 1998; Huitema 1997, as cited by Cave and Mason 2001).

True congestion delay occurs on the Internet whenever the combined traffic needing to be forwarded onto a particular outgoing link exceeds the capacity of that link. The design of the TCP assigns to the sending nodes the responsibility for regulating the flow of packets on the basis of cumulative acknowledgments from (adjacent) receiving nodes of the arrival of packets sent to them. This adaptive control mechanism operates in response to 'packet losses' that reach a rate signalling the presence of congestion to the routers that share the link. Thus, when congestion occurs, a packet may be delayed, sitting in an adjacent router's queue awaiting dispatch, and so will arrive later than some other packet from the same message that has not been subject to queuing. The result is delay in the reassembly of all the

packets that contain the message, a condition described as 'latency' in the language of telecommunications engineers. (When queue lengths vary, and some queues fill up, packets will be dropped by the router and therefore need to be resent, causing variations of the duration of delays and the condition known as 'jitter'). Congestion typically is a transient phenomenon, however, lasting only for the interval during which the TCP mechanism adapts to the available capacity by slowing the outgoing packet rate. It can reach drastic levels, however, if the capacities of the nodes available to each router fall below the minimum transmission rate provided by the control protocol.

The mechanism of congestion control provided by TCP, therefore, is simply to push back on the traffic source dynamically, in response to the detection of congestion inside the network, until it no longer is able to accept the offered load. This simple algorithm is incapable of discriminating among the initiators of the offered load, or among various types of applications that are generating traffic. Hence it cannot serve to shape the behaviour of individual users on the Internet, or even that of classes of users. Moreover, this congestion control algorithm is neither enforced on the Internet, nor is it even part of the protocol architecture of some applications that do not implement TCP - such as streaming video and UDP (User Data Protocol) (CSTB 1994: 189, 201 n. 40). Those applications consequently can be viewed as taking unfair advantage of other applications, such as email programs that do implement TCP.

Today, congestion generally is understood to be rare within the backbone networks of North American ISPs. The obvious explanation for the failure of chronic, paralyzing congestion to materialize under the conditions of unpriced usage lies in the rapid expansion of capacity to accommodate the growth of Internet hosts and traffic; and because most of the widely used applications tolerated the congestion control mechanisms provided by TCP. Whether bandwidth increases can continue to keep pace with the growth of demand, of course, depends on whether QOS-enabling enhancements are made in the network that will greatly increase the offering of bandwidth-hungry services, and the degree to which competition will either check the ability of ISPs to differentially price such services in a manner that curtails their needs for heavy investment in capacity, or result in rivalries among the larger ISPs to stake out more 'real estate on the Net in order to attract an expanded customer base.

Instead of appearing ubiquitously throughout the rest of the network, however, congestion does appear to be concentrated at particular bottlenecks created by disparities in the provision of capacity. As has been noted above, the links (exchange points) between ISPs - and especially the public network access points (NAPs) - are as a rule much more heavily congested than the links *within* the service providers' respective networks (Odlyzko 1998; CSTB 2001: 99, 117). Although the links between customers' local area networks (LANs) or residences and their ISPs are also frequently congested, the difficulty arises from the organizational delays or the expense entailed in increasing the capacity of the connection. Persistent congestion has been documented at several international links, where long and variable queuing delays, as well as high packet loss rates, have been measured (Paxson 1999; CSTB 2001: 99-100). Here again, however, the proximate source of the problem appears to be rooted in institutional circumstances affecting the provision and allocation of capacity at strategic connection points, rather than the endemic condition of unrestrained bandwidth usage envisaged by economic theorizing.

A cynical commentator might conclude that the stream of ingenious proposals from economists to fix the problem of congestion on the Internet, in typically ignoring the possible strategic explanations for congestion at the public NAPs, and proposing the introduction into the network's core of the intelligence needed to operate a sophisticated pricing mechanism, come down to the expedient of making the network less and less like the Internet, and more closely akin to a connection-oriented conventional public switched telephone network (PSTN). Quite obviously, however, had such a design been embraced to begin with, many other difficulties posed by the peculiar open-architecture would have been obviated as well. Along with removal of the obstacles to a mass transfer fee-for-service business model, this would reduce the myriad *practical* difficulties that local communities linked to the Internet now encounter in seeking to control the content of messages bearing objectionable content. In a connection-oriented system it is much more feasible to rapidly and accurately identify the locations, if not the identities of agents engaging in the electronic transmission of content that recipients deem to be pernicious - and to set about mobilizing political, social, and legal countermeasures. There would, therefore, be less need than presently exists to devote resources to the development of the still rather coarse-grained 'geo-locator technologies' that are being use to create targets for direct mail advertising and sales techniques based on the characteristics of the recipients' neighbourhoods. Less attention would also have to be given to figuring out whether such technologies can be made sufficiently reliable to be employed to control the distribution of objectionable content on the Internet, in the ways that would parallel the familiar content-regulating actions of political authorities who can identify the originating parties and have legal jurisdiction over the geographical territories in which they are situated.

Whether or not the removal of anonymity, and the reimposition of greater controls on individuals' access to content are desirable in some circumstances, is quite another matter (Engel and Keller 2000). The point is simply that the congestion-pricing solution envisaged for the Internet is not the narrow matter of economic efficiency that economists have appeared to be presenting; its implementation would require an architectural revolution in which the Internet as we know it would have disappeared. Correspondingly, in that brave new world, debates about the conflicting desiderata of privacy, anonymity, and security would continue, but they would cease to be policy matters that had a peculiar 'Internet' aspect and would simply reprise the issues that society has found ways of resolving for other communications media - physical newspapers and books, plain old telephones, radio, films, and television (de Sola Pool 1990).

3 A path to the end of 'end to end'?

Will the pressures to insert new capabilities into the core of the network really have the deleterious effects envisaged, and if undesirable consequences materialized, would it not be possible to restore the status quo ante? The likelihood is that even the unintended ending of an integral end-to-end Internet would not be readily reversible, and that the benefits thereby lost might prove difficult if not virtually impossible to recover on a later, improved successor to the global information infrastructure. This last point deserves further elaboration, which can conveniently be provided by returning to consideration of the concrete issue of permitting cable companies in the ISP market to create proprietary sub-networks on which QOS technologies are used to

offer differentiated service choices to subscribers. Users of a particular service, however, would have access only to the music and the video that their ISP had designated, possibly also to a designated IP voice telephony service, and might be similarly locked in to a particular suite of other web based services and applications software.

A concrete scenario in which this possibility might be realized is suggested by the emergence of the Skype VoIP (voice over Internet protocol) service in 2003.[3] This service is based upon the distinctive end-to-end features of the 'old' Internet and, as a consequence, free-rides on the bandwidth made available for services like the World Wide Web or e-mail. Skype's business model, which offers free voice calls between users who are connected to the Internet, is based upon charges for receiving and placing calls to users who are using other networks as well as charging for additional, premium, services such as voice mail. According to the market research company Evalueserve,[4] Skype had recruited 13 million users worldwide within two years of its founding. Sustained rapid growth of Skype will not only have a significant impact on the revenues and profitability of telecommunication network operators, but also is likely to generate significant congestion in the Internet. In addition, since Skype is a peer-to-peer technology, its use involves employing resources from individual user machines and their networks to 'relay' calls or store voice messages. This has raised corporate concerns about security and local network congestion and has led to responses such as the products offered by Bitek International which block Skype services.[5]

This situation is very similar to the past use of peer-to-peer networks to transfer music files (Skype was created by the same individuals that created Kazaa, a leading peer-to-peer file exchange application). Skype, however, involves a more complex set of issues. In some countries, for example, Israel, the use of Skype to bypass the local telecommunication operator is illegal and it is likely that congestion effects will prompt ISPs and network operators to employ blocking technologies such as those available from Bitek International. In other countries, the employment of blocking or filtering technologies is likely to be more decentralized and involve corporate networks or specific ISPs. These developments are likely to lead to a growth in 'restricted web' services, that is, those that utilize some applications and block others. It is a short step from these developments to legitimizing Skype by incorporating it as a specific service offered by leading ISPs. Skype users would thus be assured (in countries permitting such services) that they would be able to retain connectivity with other Skype users as well as telephone services to connect to others.

There are reasons to expect that having put in place a restricted-web ISP service offering, such as Skype, the ISP in question might well be receptive to allowing compatibility between this sub-network and other similar subnetworks. The economic logic of this situation differs from that which governs in the general analysis of compatibility standardization for network interoperability - where it is generally found that small networks seek connectivity with larger ones, and the latter have stronger incentives to remain aloof from rivals of comparable size.[6] By linking with similarly sized networks, an ISP with a large network base could offer subscribers other enhanced services that, like voice telephony, are latency-sensitive, such as multiple-player interactive video games, and a larger choice among the set of preselected applications. The value of integrating to achieve compatibility with smaller ISPs would remain comparatively small, and so, in this market setting, the dynamics lead toward a high degree of market power concentrated in the hands of a

small number of ISPs, and a large fringe of ISPs whose clientele remains cut off from these enhanced services.

Thus entrenched, the dominant ISPs would be in a position to extract some if not most of the rent that might otherwise flow to the developers of applications innovations, in exchange for making these available for use by their clientele. Lacking that access, the developers would be confined to exploiting niche markets at the fringes of the network, where their products would remain beyond the reach of the subscribers to the large ISPs. Nothing in this picture suggests that the emergent structure of a partitioned network would be likely to be voluntarily dismantled by the incumbent, or vertically integrated ISPs, nor successfully attacked by an entrant possessing a novel and superior application technology. An entrant with the capital resources required to establish a new, competitive, vertically integrated ISP, moreover, would have every incentive to seek compatibility with an existing large service provider, and an aggressive newcomer aggressive might expand by stealing the original incumbent's clientele. But, in addition to requiring the financial backing to create the additional network capacity required for the implementation of that strategy, the successful entrant would replicate the initial situation, and pose an even greater entry barrier to the next innovator.

A mitigating consideration to be noted in the foregoing scenario is that although the technological enhancements to the Internet would create new opportunities for ISPs to extract greater rents (consumer surplus) from their customers by means of discriminatory pricing schemes (Mandjes 2004; Odlyzko 2004), the strategy of vertical bundling of networking services and Internet-based applications, nevertheless, would provide additional benefits for a large segment of the Internet population. The technologists who created an end-to-end architecture, and who value it particularly for the support it provided to applications innovators, are less burdened than the typical Internet user by having to install, configure, upgrade, and maintain the software of each and every one of the rapidly growing number of applications that must be attached at the networks' end points. This state of affairs can be expected only to become more burdensome. As Blumenthal and Clark (2001: 72) perceptively observe:

> "The importance of ease of use will only grow with the changing nature of consumer computing. The computing world today includes more than PCs. It has embedded processors, portable user interface devices such as computing appliances or personal digital assistants (PDAs, such as Palm devices), Web-enabled television and advanced set-top boxes, new kinds of cell-phones, and so on. If the consumer is required to set up and configure separately each networked device he [sic!] owns, what is the chance that at least one of them will be configured incorrectly. That risk would be lower with delegation of configuration, protection, and control to a common point, which can act as an agent for a pool of devices. This common point would become a part of the application execution context there would no longer be a single indivisible end-point where the application runs."

While pointing to the threat to the preservation of the open-network architecture, this acknowledges that the creation by ISPs of enclaves containing advanced services would be one way in which the multitude of less technically sophisticated users could obtain specialized (and correspondingly standardized)

network applications-integrating services. Thus, in regard to this issue -- as is the case in so many others, network policymakers face the classic tradeoff of securing the immediate benefits of closed standardization by sacrificing the technological flexibility that is conducive to future radical innovations (David 1995).

4 Policy priorities and protection of the Internet's architecture

It has been seen that among the many technological fixes proposed for enhancing the Internet's performance some are not so innocuous, because they would entail inserting intelligence into the core of the network. The likely impact of these induced innovations therefore would be the alteration of the distinctive end-to-end architecture, pushing the future path of the network's evolution more towards emulating the performance features (both good and bad) associated with a connection-oriented telecommunications system - the familiar paradigm of which exists in the PSTN (David and Steinmueller 1996). Will the changing balance among the interests of the communities using the information infrastructure, inevitably force a sacrifice of the global infrastructure's transparency and openness, thereby raising new barriers to the invention and diffusion of valuable applications? Inasmuch as a technological drift away from the original Internet's end-to-end architectural design should not be regarded as an inexorable process beyond the reach of social control, there is scope for policy interventions to check such a course of evolution. It must be hoped, then, that promoting wider understanding of the issues at stake can increase the political feasibility of arriving at rational policy priorities. At least, that is the spirit in which the following commentary on the identification and balancing among conflicting goods will proceed.

A first appropriate step is to ask whether the net impact of any proposed movement in that direction would be socially beneficial. In view of the prospective emergence of a broadband Internet on which QOS will be more widely implemented by ISPs competing for customers while seeking the means to charge what the (multimedia) traffic will bear, the question might be asked whether the time has come for end-to-end to end. It could be argued that inasmuch as the days of Internet1 as a unified global infrastructure providing a receptive platform for rapid innovation and experimentation with networks are numbered, the best course of action would be to make whatever changes are required in the core of the network to quickly reap the benefits of the available new services on a "users' Internet.'' That is to say, we should come to terms with the immanent tendency of the evolutionary dynamics driven by the needs of the maturing market for a differentiated Internet service, and think about other ways to provide a network environment that would stimulate the continuation of amazing innovations.

Such a view would counsel turning attention to the construction of a separate, very high speed internetwork as the test bed and experimental commercial market for advanced services, which would be designed to provide the features of openness and flexibility that have proved so encouraging to the development of more powerful digital technologies. This might be called Internet2+ to distinguish it from the actual federally funded backbone created in the U.S. to continue the National Science Foundation's Network (NSFNET) research role. There is something to be said for this vision of a cyclical regeneration of a new inter-networking environment that would revive some characteristics of the original. It acknowledges the important symbiotic relationship between the mature PSTN infrastructure on which packet switching and

the novel technologies of the ARPANET (Advanced Research Projects Administration Network) and NSFNET could be erected; and it recognizes the fertility of experimental research communities as sources of user-designed technological innovations. But, unfortunately, it ignores the crucial fact that an important aspect of the historical experience cannot be replicated or revived by these means.

The nub of the problem is that to develop innovations that are readily available for deployment on the Internet as it exists, one needs a test bed with its technical features. Yet, for the communities that would have access to Internet2+, and especially for those groups that are engaged in advancing the frontiers of network engineering, the high value use would be to develop applications that utilized the enhanced properties of that infrastructure rather than the more limited capabilities of Internet1 - or the still less accommodating infrastructure into which the latter would be tending to evolve.

Next, one should consider the net balance of gains against losses: would the contemplated enhancements in the quality of differentiated services, and in the ability of service providers to engage in price discrimination among the users of the Internet, compensate for the economic welfare costs entailed - in terms of curtailed future scaleability and a slowed pace of innovation in applications? Several grounds for scepticism regarding the value of the gains seem worth keeping in mind.

To begin with, the incremental social benefit of upgrading the Internet to carry real-time audio traffic is not obviously overwhelming, given the existence of other technological means of providing a large part of the world's population with access to voice telephony (via cellular radio and satellite transmission) at lower fixed costs than those entailed in laying copper wire or fibre-optic cabling. Certainly, Internet telephony could be integrated into new, multimedia services. Yet, there is a disjunction here between a strategy directed toward opening profit opportunities in the developed economies - to elicit continued private sector investment in augmenting the broadband infrastructure available to users in those countries - and a policy that also takes account of the situation in the world at large.

While cellphone technology has opened the benefits of rapid, global communications to large cohorts in the developing economies, it remains unsuitable for sparsely populated regions and geographically remote sites, just as it is not capable of supporting the very high bandwidth communications that are likely eventually to be in demand there. But systems of low earth orbit satellites (LEOS), which are designed to provide two-way, low-latency, point-to-point transmissions, will be available to fill these significant service gaps. According to expert engineering opinion, the seamless linking of LEO satellite constellations into the worldwide communications infrastructure is a development that can be expected to take place in the relatively near future.[7]

For the developing economies, however, it is accepted that even to provide substantial narrowband coverage, considerable amounts of public funding for upgrading existing telecommunications infrastructures would be necessary; and some of that is likely to be provided by subsidized loans and transfers through multinational cooperative agencies. It must therefore be recognized that the social rate of return on public (and private) investments in this infrastructure would be reduced substantially if the present core of the Internet were to be modified by engineering changes that deviated from the principles of end-to-end. To permit alterations to the architecture of

the backbone networks in the high income countries, in order to provide users there with Internet voice telephony (along with business or entertainment services integrating real-time video), would effectively mean curtailing the access afforded newly connected users in the world's poorer societies to existing information tools and global data resources.

The foregoing remarks address possible discrepancies between the private incentives driving the Internet's technological evolution, and the social value of the enhancements that would thus be achieved. They have not touched on the need to explore engineering improvements that can be implemented (at the edges of the network) in ways that would not compromise the performance attributes that derived from the Internet's end-to-end architecture. Content labelling conventions are one example of the kind of "improvements" that, if voluntary adopted or enforced on content providers, would enhance the efficiency of filtering at the endpoints of the network.

But another important set of alternatives to introducing control mechanisms in the network's core that remains to be considered is the large class of *non*-technological options. In view of the fact that the origins of many of the vexing dysfunctionalities of the Internet derive from the historical displacement of this technology system from the peculiar, highly regulated behavioural and organizational contexts within which it was created and initially used, an obvious option to be considered is the restoration of some of the former modes of regulating users' behaviours. The Internet may have been a technology that quite by accident was well-attuned to the laissez-faire spirit of the era in which it was publicly introduced. Yet, an ideologically driven commitment to go on thinking exclusively in the same vein about ways to overcome the problems posed by the 'network of networks', rejecting social engineering in favour of solutions found through Internet reengineering, is most likely to sacrifice the Internet's unique and valuable pro-innovation features. There is no a priori reason to conclude that the most efficient solution path is one that relies solely on fixes that can be technologically implemented.

Yet, proposed regulation and interventions by public authorities continue to be opposed on the argument that such actions are inimical to the Internet's survival as a global interaction space free from governmentally imposed structures of social regulation. Current rhetorical support for relying on engineers to fix whatever might really need mending, rather than letting legislators and lawyers loose in cyberspace, presents a curious mixture of attitudes. These are compounded from the libertarian philosophy that is pervasive among survivors of the Internet's pioneering user groups, strains of anarchosyndicalism that have emerged in the ethos of the latter-day hacker culture, and the generic laissez-faire disposition of the Internet's more recently arrived community of business entrepreneurs. The holders of pro-commercial and anti-commercial sentiments alike appear quite comfortable making common cause against the intrusion of government regulations that are socially engineered. This, it should be recognized, presents an essential political and philosophical position, quite distinct from the utilitarian rationale that would give priority to preserving the distinctive end-to-end architecture of the Internet - especially inasmuch as serving the latter priority might call for the development of new, institutional mechanisms of governance.

Lawyers looking at the evolving Internet are naturally disposed to pose this issue in terms of a political choice between the regulation of human actions by laws

or governance by 'Code' - the encompassing term used by Lessig (1999) in referring to the architectural configuration of networks and the location of access points, the design of hardware, operating systems, languages, data formats, and applications software. Economists, it would seem, would have something helpful to contribute to debates on these questions, by directing attention to the relative costs of alternative modes of regulation in network environments, especially in view of the significant externalities and irreversibilities that are likely to be entailed by introducing either technological or institutional modifications in the existing regime (see Mueller 2002).

Furthermore, approaching some questions that involve the governance of human behaviour in cyberspace from the perspective of the economics of crime and punishment also may be a useful way to mediate in debates between the engineers and the lawyers: the quest for perfect technological mechanisms of detection and suppression of malefactors is only relevant in a perfect world, and it is possible to compensate for reduced probabilities of being caught by raising the penalties visited on those who are. This approach may not be good enough in some areas of concern, and other technological safeguards will be needed to protect humans and vital technological systems alike from grave damage. But much of the 'protective' control of behaviour afforded under the law has been found to work tolerably well with this mixed approach.

For those reasons and still others, the relevant policy questions ought not to be construed in terms of making either/or choices. It is important to resist the rhetoric of much contemporary discussion of economic policy, which tends to offer only extreme alternatives. Participants are too often driven into opposing camps, one side calling for the introduction of government controls, and the other placing its faith on the further development of decentralized, automatic, supposedly neutral, and (market-like) regulatory mechanisms that can better resist political manipulation and so preserve greater scope for human volition. The following statement exemplifies the polarizing impact of applying the technologists' Internet philosophy to decide on the best means of protecting privacy on the Net:

> "[T]he cyperpunk credo can be roughly paraphrased as 'privacy through technology, not through legislation'. If we can guarantee privacy protection through the laws of mathematics rather than the laws of men and whims of bureaucrats, then we will have made an important contribution to society. It is this vision which guides and motivates our approach to Internet privacy." (Goldberg, Wagner, and Brewer. 1997, quoted in Blumenthal and Clark 2001: 84 n. 52)

A full-blown systems design approach, by contrast, would hold that if the benefits of the Internet's end-to-end architecture are to be retained, some technological solutions simply cannot be substituted for other, socio-legal modes of governing the behaviour of agents on the Internet. Rather than being viewed as antithetical substitutes, the potential complementary of technological and institutional mechanisms governing the digital communications infrastructure should be explored in a coordinated manner.

There is thus a case to be made for devoting greater attention to matching the technological innovations of the Internet by mobilizing other, nontechnologically implemented modes of regulation. Greater consideration surely is worth directing to

the design of legal, political, and social rule structures and administrative procedures, of the kind that proved to be efficacious in supporting successful economic exploitation of previous technical advances in communications networks. In this connection it is worth recalling that the oldest international treaty organization in existence today is the International Telecommunications Union (ITU). This institution, which began its life in 1865 as the International Telegraph Union, provided the model in whose image virtually all subsequent international treaty-based organizations were created (David and Schurmer 1996; Schmidt and Werle 1998). While that may suffice to suggest the possibility that fruitful innovations in international rule-making fora can be driven by the opportunities, or problems, that new technologies create, there is no doubt that today very formidable challenges are posed for the adaptive coevolution of international laws governing cyberspace (Gamble 1999.)

Conclusion

Even as the Internet comes of age, the technology of the global information infrastructure and the organization of the communication service industries based on it continue to undergo significant changes. The main message carried by the foregoing discussion is that many microeconomic policy recommendations and engineering proposals that have been presented as incremental modifications to enhance the performance capabilities of the Internet actually may have radical implications for the future course of its technological evolution. These have been seen to involve rather esoteric matters that might appear best left to be decided by engineering specialists, and experts in the intricacies of telecommunications regulations. But decisions taken in those realms will powerfully shape the future performance characteristics of the Internet. In that way, they will have important consequences for the nature, size, and distribution of the economic and social benefits that it yields.

It is understandable that the initial reaction of many economists who had developed familiarity and expertise in the context of studying mature telecommunications networks (that is, the PSTN) found it natural to transfer to the sphere of Internet economics the modes of analyses and policy prescriptions that were, so to speak, most ready to hand. Thus, a great deal of prominence has been given to the discussion of principles that should govern optimal pricing of access to the transport/bearer layer of the Internet, a matter of undoubted importance for existing and would-be service providers. In a technologically dynamic network setting such as that of the Internet, however, the feasibility and terms of entry also depend on nonprice policies, including those affecting technical compatibility standards, and regulations governing the interconnection strategies of incumbent service providers (Cave and Mason 2001). Over the long run, the technical rules of the game affecting physical interconnection are likely to be more consequential than pricing formulae in their effects on the growth and distribution of available bandwidth, competition in the ISP market, and the rate of innovation in applications on the Internet.

Bertrand Russell once remarked that we must ‘tolerate specialists because they do good work.’ Perhaps it would be more generous to speak of appreciation rather than toleration, but the point remains that in matters whose potential implications for human welfare are as important as those at hand, more than narrow expertise is

wanted. The story of the Internet's development justly can be presented as a remarkable case of 'success by design' (CSTB (2001: 34) invokes this phase in discussing architectural principles). Equally, it may be read as a path-dependent tale of fortuitous engineering design decisions that were made with little consideration for aspects that have turned out to be problematic for many of the purposes and social contexts in which the resultant, wonderfully open and flexible technology would be used.[8]

As societies around the world continue to wrestle with difficult technical challenges and policy quandaries that have their origins in historically remote decisions that proved to be essentially irreversible, an obvious question to be asked is whether it has become possible now to proceed differently. The historical economics approach (David 2001b) that informs much of the foregoing discussion carries some additional and potentially more provocative suggestions for rethinking the economics of the telecommunications regulation in the age of the Internet. Because economic analysis of industrial organization and public regulation of telecommunications utilities was developed with reference to industries based on a mature network technology, practitioners in this area remain too inclined to start from the assumption that the technology is given.

This is seldom the case, and it is palpably misleading when applied to the situation of the Internet. Therefore, perhaps the most important general lesson to be drawn for the future of Internet policy analysis is for economists to start thinking about the ways in which the structure of the existing markets, and the uneven and uncoordinated regime of regulation and nonregulation, may induce research and technological innovation to take some directions while discouraging technical progress from proceeding in others.

Acknowledgements

In this contribution I have drawn upon previously published writings of mine that benefited from the informative comments of Marjory Blumenthal, Andrew Glyn, Andrew Graham, Robert Spinrad, Gregory Rosston and Raymund Werle. I am particularly grateful to Robin Mansell and W. Edward Steinmueller for their substantive and editorial improvements upon an earlier draft of this chapter, but responsibility for errors, omissions and the sometime captious views expressed remains mine alone.

References

Blumenthal, M. S. and Clark, D. D. (2001). 'Rethinking and Design of the Internet: The End to End Argument vs. the Brave New World'. *ACM Transactions on Internet Technology,* 1(1): 70-109.

Cannon, R. (2005) 'State Regulatory Approaches to VOIP: Policy, Implementation, and Outcome', *Federal Communications Law Journal*, 57(3): 479-510.

Cave, M., and Mason, R. (2001). 'The Economics and Regulation of the Internet'. *Oxford Review of Economic Policy*, 17(2): 188-201.

CSTB (Computer Science and Telecommunications Board) (1994). *Realizing the Information Future: The Internet and Beyond*. Washington, DC: CSTB.

——(2001). 'Computer Science and Telecommunications Board, National Research Council'. *The Internet's Coming of Age*, Washington, DC: National Academy Press.

David, P. A. (2001a). 'The Evolving Accidental Information Super-highway'. *Oxford Review of Economic Policy*, 17(2): 159-87.

———(2001b). 'Path Dependence, Its Critics and the Quest for "Historical Economics', in P. Garrouste and S. Ionnides (eds) *Evolution and Path Dependence in Economic Ideas: Past and Present.* Cheltenham: Edward Elgar Publishing, 15-40.

———(1995). 'Standardization Policies for Network Technologies: The Flux Between Freedom and Order Revisited,' in R. Hawkins, R. Mansell, and J. Skea (eds) *Standards, Innovation and Competitiveness: The Political Economy of Standards in Natural and Technological Environments.* Cheltenham: Edward Elgar, 15–35.

——and Greenstein, S. (1990). 'The Economics of Compatibility Standards: A Review of Recent Research'. *Economics of Innovation and New Technology*, 1(1&2): 3-41.

——(with Shurmer, M.) (1996). 'Formal Standards-Setting for Global Telecommunication and Information Services'. *Telecommunications Policy*, 20(10): 789-815.

——and Steinmueller, W. E. (1996). 'Standards, Trade and Competition in the Emerging Global Information Infrastructure Environment'. *Telecommunications Policy*, 20(10): 817-30.

de Sola Pool, I. (1990). *Technologies Without Boundaries: On Telecommunications in a Global Age*. Cambridge, MA and London: Harvard University Press.

Engel, C. and Keller, K.H. (eds) (2000). *Understanding the Impact of Global Networks on Local Social, Political and Cultural Values.* Baden-Baden: Nomos Verlagsgesellschaft.

Farrell, J. and Saloner, G. (1986). 'Installed Base and Compatibility: Innovation, Product Preannouncements and Predation'. *American Economic Review*, 76(4): 940-55.

Gamble, J. K. (1999). 'New Information Technologies and the Sources of International Law: Convergence, Divergence, Obsolescence and/or

Transformation'. *German Yearbook of International Law*, Berlin: Duncker and Humbolt, 170-205.

Goldberg, I., Wagner, D., and Brewer, E. (1997). 'Privacy-enhancing Technologies for the Internet', University of California, Berkeley, 42nd IEEE COMPCOM'97 Conference, Spring, http://www.cs.berkeley.edu/~daw/papers/privacy-compcon97-www/privacy-html.html, accessed 16 Mar. 06.

Graham, A. (2001). 'The Assessment',Introduction to Special Issue on the Economics of the Internet. *Oxford Review of Economic Policy*, 17(2): 145-58.

Gupte, R. P. (2001). 'Pricing to Control Congestion: An Economist's Bias', Trinity Term Research Paper in Economics 168X, Stanford University Centre in Oxford, June.

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

Lessig, L. (1999). *Code and Other Laws of Cyberspace.* New York: Basic Books.

MacKie-Mason, J. K. and Varian, H. R. (1995a). 'Pricing the Internet' in B. Kahin and J. Keller (eds), *Public Access to the Internet*. Cambridge, MA: MIT Press, 269-314.

——.(1995b). 'Pricing Congestible Network Resources'. *IEEE Journal of Selected Areas in Communications*, 13(7): 1141-9.

——— (1997). 'Economic FAQs About the Internet', in L. W. McKnight and J. P. Bailey (eds) *Internet Economics*. Cambridge, MA: MIT Press, 27-62.

MacKie-Mason, J. K. and Varian, H. R. (1996). 'Some Economics of the Internet,' in W. Sichel (ed.), *Networks, Infrastructure and the New Task for Regulation*. Ann Arbor MI: University of Michigan Press, 107-36.

Mandjes, M. (2004) 'Pricing Strategies and Service Differentiation', *Netnomics*, 6(1): 59-81.

McKnight, L.W. and Bailey, J. P. (1997). 'An Introduction to Internet Economics,' in L. W. McKnight and J. P. Bailey (eds), *Internet Economics*. Cambridge, MA: MIT Press, 3-26.

Mueller, M. (2002) *Ruling the Root: Internet Governance and the Taming of Cyberspace*. Cambridge MA: MIT Press.

Odlyzko, A. (1998). 'The Economics of the Internet: Utility, Utilization, Pricing and Quality of Service', AT & T Labs-Research, http://www.dtc.umn.edu/~odlyzko/doc/internet.economics.pdf , accessed 17 Mar. 06.

Odlyzko, A. (2004). 'Privacy, Economics, and Price Discrimination on the Internet', in L. J. Camp and S. Lewis (eds) *Economics of Information Security - Advances in Information Security, Vol. 12*. Boston, Dordrecht and London: Kluwer Academic, 187-211.

OECD (2006). 'Policy Considerations of VOIP', Working Party on Telecommunication and Information Services Policies', DSTI/ICCP/TISP(2005)13/Final, Paris, 20 March.

Paxson, V. (1999). 'End-to-End Internet Packet Dynamics'. *IEEE/ACM Transactions on Networking*, 7(3): 277-92.

Schmidt, S. K. andWerle, R. (1998). *Coordinating Technology. Studies in the International Standardization of Telecommunications.* Cambridge MA: The MIT Press.

Shapiro, C. and Varian, H. R. (1999). *Information Rules: A Strategic Guide to the Network Economy*. Boston, MA: Harvard Business School Press.

Varian, H. R., Farrell, J. and Shapiro, C. (2004). *The Economics of Information Technology: An Introduction*. Cambridge: Cambridge University Press.

Endnotes

[1] See http://www.tprc.org/ARCHIVES.HTM , accessed 17 March, 2006.

[2] For a representative example, drawn more or less at random from online listings, see, Prof. P. K. Dutta's course lectures: http://www.columbia.edu/~pkd1/lecture20002.html, accessed March 10, 2006.

[3] For a discussion of policy and regulatory approaches to VoIP in the US, see Cannon (2005) and more generally, OECD (2006).

[4] See http://www.evalueserve.com/, accessed 17 March, 2006.

[5] See http://www.bitek.com/, accessed 17 March, 2006.

[6] Farrell and Saloner 1986; David and Greenstein 1990, for a review of the early literature; Shapiro and Varian 1999, especially chapters 7 and 9 on strategies in standards wars; and Varian, Farrell and Shapiro 2004.

[7] Private communication from Robert Spinrad, 9 May 2000.

[8] On concepts of irreversibility, path-dependence, and 'path-constrained melioration', see David 2001b.