Granting network providers pricing flexibility should reduce the costs borne by consumers.

Network Neutrality or Internet Innovation?

By Christopher S. Yoo
University of Pennsylvania Law School

Network neutrality has received sustained attention from both policymakers and academic commentators for the past several years, and it shows no signs of retreating from the forefront of the policy debate. President Obama effectively ensured that network neutrality will remain at the top of the policy agenda by including provisions in the 2009 stimulus package that require the Federal Communications Commission to formulate a national broadband plan. The stimulus package also requires that grants made by the National Telecommunications and Information Administration comply with four network neutrality principles first articulated by the FCC in 2005. On October 22, 2009, the FCC initiated proceedings to codify and expand the 2005 principles. President Obama reaffirmed his support for network neutrality in a YouTube interview conducted shortly after his 2010 State of the Union address.

Pinning down a precise definition of network neutrality is difficult. Roughly speaking, it requires network providers to route traffic without regard to the source or content of the packets of data that move across the Internet, the application with which those packets are associated, or the sender's willingness to pay. In the words of leading network neutrality proponent Lawrence Lessig, "Net neutrality means simply that all like Internet content must be treated alike and move at the same speed over the network."

It would be surprising if any two similar packets would be treated exactly alike when traveling through a network consisting of more than 30,000 autonomous systems that determine their terms of interconnection through arms-length negotiations. Indeed, many commentators have noted that such equal treatment did not occur over much of the Internet's past, when it was far less complex. Now, systematic changes in the architecture of the Internet make identical treatment even less likely, yet the changes are largely the result of network providers' attempts to reduce cost, manage congestion, and maintain quality of service. These changes may not represent network providers' efforts to promote their self interests at the expense of the public, as some network neutrality proponents have suggested, but instead they have the potential to yield substantial benefits both to individual consumers and to society as a whole.

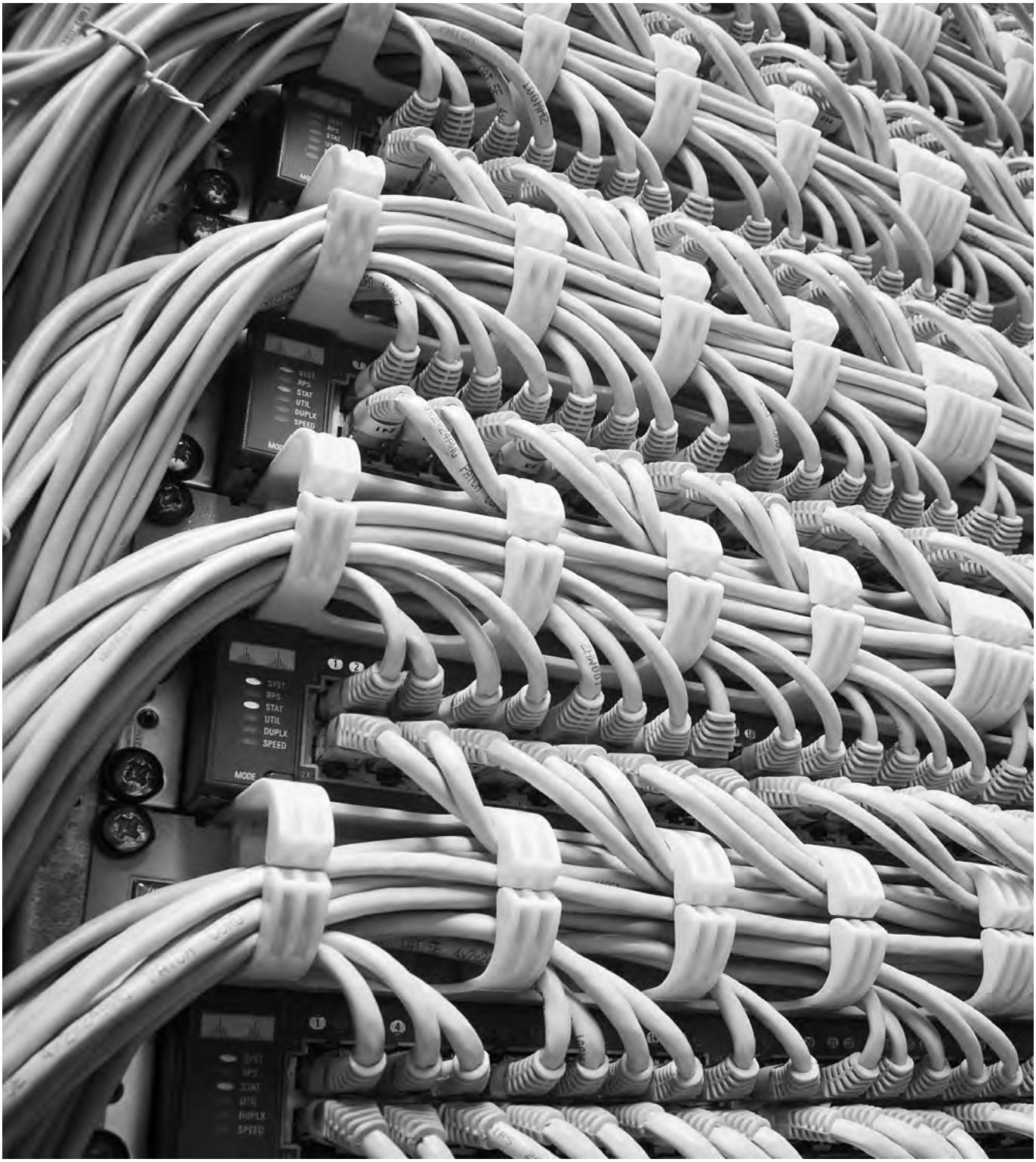
THE EARLY INTERNET

When the Internet first emerged, its topology and the business relationships comprising it were relatively simple. The Internet evolved out of the National Science Foundation's NSFNET backbone, which was created in 1986 (and decommissioned in 1997) to provide universities all over the country with access to federally funded supercomputing centers located at five major universities. The primary architects of NSFNET decided to give it a tripartite structure. At the top was the NSFNET backbone, which at its peak connected 16 research facilities across the country. At the bottom were the campus networks run by individual universities. In the middle were regional networks (typically operated by university consortia or state-university partnerships) that linked the campus networks to the major computing centers.

Every data packet had to travel through a parallel path traversing each level of the hierarchy. For example, traffic originating on one campus network would have to connect to the regional network with which it was associated, which hand-

Christopher S. Yoo is professor of law and communication at the University of Pennsylvania and founding director of Penn's Center for Technology, Innovation, and Competition.

This article is adapted from "Innovations in the Internet's Architecture that Challenge the Status Quo," appearing in the Winter 2010 issue of the *Journal on Telecommunications and High Technology Law*.



MAXIM FILIPCHUK/ISTOCKPHOTO

ed off the traffic to the NSFNET backbone, which in turn handed it off to the regional network that served the destination campus network. The result was to create a series of parallel hierarchies through which all traffic had to traverse.

The network retained this same basic architecture when it was privatized during the mid-1990s. The NSFNET backbone at the top of the hierarchy was replaced by a series of private backbone providers that interconnected with one another at four public network access points established by the National Science Foundation. The campus networks at the bottom of the hierarchy were replaced by last-mile providers that transported traffic from local distribution facilities located in individual cities (which in the case of digital subscriber lines are usually called "central offices" and in the case of cable modem systems are usually called "headend") to end users' residences and places of business. The regional networks evolved into regional Internet service providers (ISPs) that transported traffic between the four network access points served by backbone providers and the central offices and headends maintained by last-mile providers.

The privatization of the Internet did not change the hierarchical nature of the basic architecture. Each regional ISP still connected to a single backbone provider, and each last-mile provider still connected to a single regional ISP. Indeed, the early versions of the routing protocol employed by the backbones (known as "border gateway protocol") would not support more complex topologies.

This architecture conferred a number of advantages. It constituted a "spanning tree" that connected all of the nodes with the minimum number of links. Furthermore, the fact that the path between any two nodes was unique greatly simplified determining the path along which traffic should be routed. That said, tree architectures are also subject to a number of drawbacks. The uniqueness of the path connecting any two nodes means that the failure of any link or node in the network will inevitably disconnect part of the network. Even when all network elements are operating properly, if the rate at which traffic arrives exceeds any particular element's capacity to route the traffic, that network element will become congested and the quality of service provided will deteriorate. In addition, the hierarchical structure made each network participant completely dependent on the players operating at the level above them, which in turn provided backbones with a potential source of market power.

Peering and Transit The early Internet was also characterized by relatively simple business relationships. End users typically purchased Internet access through some form of "all-you-can-eat" pricing, which allowed them to consume as much bandwidth as they would like for a single flat rate. Relationships between network providers typically fell into two categories. Tier-1 ISPs entered into "peering" relationships with one another, in which they exchanged traffic on a settlement-free basis and no money changed hands. The primary justification for foregoing payment is transaction costs. Although the backbones could meter and bill each other for the traffic they exchanged, they could avoid the cost of doing so without suffering any economic harm so long as the traffic they exchanged was roughly symmetrical; such arrangements would not be economical if the traffic being exchanged were severely imbalanced. Thus tier-1 ISPs will not peer with other networks that are unable to maintain a minimum level of traffic volume. In addition, peering partners typically require that inbound and outbound traffic not exceed a certain ratio. Networks that cannot meet these requirements must enter into "transit" arrangements in which they pay the backbone to provide connectivity to the rest of the Internet.

Most early analyses of these arrangements focused on their financial terms. What is often overlooked is that interconnection agreements covered two distinct functions: the sending and receiving of traffic, and the announcing to the rest of the Internet where IP addresses served by various providers are located. To understand this latter function, consider the perspective of a small network, *A*, that serves a small number of its own customers and purchases access to the rest of the Internet through another ISP. The transit agreement between *A* and the ISP would not only require the ISP to receive traffic sent by *A* and to deliver traffic bound to *A*, but also require the ISP to announce to the rest of the Internet how to reach the IP prefixes associated with *A*'s customers. In addition, *A* can maintain a very simple routing table — it need only keep track of the prefixes of the customers that it serves; for all IP addresses outside of *A*, it can enter a "default route" into its routing table that directs all other traffic to the other ISP.

The existence of default routes creates a potential problem. If none of the routing tables involved in a particular routing session contained the location of the destination, by default the networks would simply hand the packets back and forth continuously and the packets would never reach their final destination. The only way to avoid this problem is for one or more network providers to maintain routing tables that map the entire Internet without employing any default routes. Thus, tier-1 ISPs are defined not only by their engaging in settlement-free peering with one another, but also by their maintaining routing tables that contain no defaults. Peering contracts also include a number of other requirements to guard against free riding and to ensure the proper functioning of the network.

THE INTERNET'S EVOLUTION

Over the past decade, ISPs have begun to enter into more complex interconnection arrangements that deviate from the strict tripartite hierarchy that characterized the early Internet. In addition, content providers have begun to experiment with a variety of ways to locate their content closer to end users. Both types of changes have significant implications that have largely been overlooked in the policy debate.

Private Peering, Multihoming, and Secondary Peering One of the first problems to emerge in the early Internet was congestion at the four network access points, which often caused throughput times and network reliability to degrade. Some estimate that this congestion caused packet loss at rates as high as 40 percent. As the network access points became increasingly congested, backbones began to find it advantageous to exchange traffic at private interconnection points, a practice known as "private peering."

In addition, regional ISPs have begun to connect to more than one backbone, a practice known as "multihoming," in part to protect against service outages and to limit their vulnerability to any exertion of market power by a backbone. Regional ISPs that did not have sufficient volume to peer with the tier-1 backbones also began to find that they did have sufficient volume to peer with other regional ISPs, a practice known as "secondary peering." Enabling regional ISPs to exchange traffic on a settlement-free basis reduced the costs borne by end users. In addition, secondary peering would often shorten the number of hops needed for particular packets to reach their final destination and make them subject to bilateral (as opposed to multiparty) negotiations, both of which should increase networks' control over quality of service. Secondary peering and multihoming also made the network more robust by creating multiple paths through

which network nodes could interconnect. In fact, as much as 70 percent of the nodes in the Internet can now communicate with one another without passing through the public backbone. This had the additional benefit of weakening the market position of the top-tier backbones, since any breakdown in the business relationship would not necessarily disconnect the ISP from the network and the ability to route along different paths places a natural limit on the backbones' ability to engage in supracompetitive pricing.

The emergence of interconnection relationships that deviate from the strict hierarchy that characterized the early Internet represents a substantial divergence from network neutrality. For example, assume that an end user is downloading content from both CNN.com and MSNBC.com. Assume further that the end user's regional ISP has a secondary peering relationship with the regional ISP serving CNN.com, but does not have a secondary peering relationship with the regional ISP serving MSNBC.com. The absence of a secondary peering relationship means that traffic from MSNBC.com will have to pay transit charges, while traffic from CNN.com will not. The result is that traffic that is functionally identical will end up paying different amounts. The differences in topology may also allow the traffic from CNN.com to maintain greater control over the quality of service.

The presence of multiple routes between these two points also complicates routing decisions. The presence of multiple paths connecting two points naturally means that someone must decide along which path to route the traffic. Although most networks choose routes that minimize the number of hops, networks may sometimes find it beneficial to route traffic in order to satisfy other requirements of their interconnection relationships. For example, a network may seek to enhance efficiency by balancing the loads between the two links. Multihomed entities can also monitor the quality of service provided by each connection and route the most delay-sensitive traffic along the link with the lowest latency.

In addition, transit contracts call for customers to pay a flat fee up to a predetermined peak volume (known as the committed rate) and pay additional charges for any volume that exceeds that level. For the same reason that consumers with two mobile telephones have the incentive to use up all of the prepaid minutes on both lines before incurring any additional per-minute charges, multihomed entities have the incentive to utilize all of their committed rate before paying additional fees. This lowers overall transit cost, but requires diverting some traffic along a path that is longer than the one stored in the routing tables. For similar reasons, a network may intentionally route traffic over a more costly path if doing so will help it maintain its traffic within the ratios mandated by its peering contract. Again, the effect is to introduce significant variance in the speed with which similarly situated packets will arrive at their destination and the cost that similarly situated packets will have to bear. This variance results not from anticompetitive motives, but rather from networks' attempts to minimize costs and ensure quality of service in the face of a network topology that is increasingly heterogeneous.

Server Farms and CDNS Large content providers have begun to employ other means to reduce cost and manage latency. One solution is to forgo maintaining a single large server and instead to deploy multiple points of presence in "carrier hotels" across the country. Doing so allows these content providers to avoid paying transit charges to reach the public backbone and instead transmit their traffic through secondary peering arrangements with tier-2 ISPs. Greater reliance on private networks also gives the content providers greater control over network security and performance. A recent study indicates that Google, Yahoo!, and Microsoft have been able to use server farms to bypass the backbone altogether for roughly a third of their traffic, and to keep their number of hops for traffic that had to pass through the backbone to no more than one or two.

On other occasions, content providers are distributing their data through "content delivery networks" (CDNs) such as Akamai and Limelight. CDNs in effect substitute storage for long-distance networking capacity by maintaining a network of local caches across the Internet. When an end user sends a request for a webpage hosted by a CDN, that query is redirected to the cache. CDNs are thus able to use storage to serve multiple queries for the same content without using significant network resources. The geographic dispersion of the caches usually dictates that the file will be served by a location closer than would be possible if all of the content were stored in a central server, which minimizes cost and latency. The distributed nature of the caches also provides protection against denial-of-service attacks and allows the CDN to redirect queries to other caches when particular caches are overly congested.

CDNs represent an innovative way to deal with the increasing complexity of the Internet. The problem is that they are nonneutral. CDNs work best for static content; they are less well suited to interactive content that changes dynamically. More to the point, CDNs are commercial services; thus greater

Secondary peering and multihoming have the benefit of weakening the market position of the top-tier backbones.

reliability and quality of service are available only to those who are willing to pay for them.

To the extent that CDNs use the public backbone to deliver the content to their caches, they are best regarded as an overlay to the existing network. Increasingly, however, CDNs and server farms are bypassing the public backbone altogether and connecting to their caches through private networks, in the process transforming CDNs into a fundamentally different architecture.

All of these developments represent innovative adjustments to the realities of the Internet. The differences in topology mean that traffic that is otherwise similar may travel through the network at different speeds, with different costs, and with different levels of quality of service.

THE EVOLUTION OF BUSINESS RELATIONSHIPS

The evolution of the Internet has not been restricted to topology. Network participants have also been experimenting with an increasingly broad range of business arrangements. Some of these innovations have been driven by the increasing significance of peer-to-peer technologies. Other important developments are partial transit and paid peering.

Peer-To-Peer One of the primary forces causing business relationships to change is the growing importance of applications using peer-to-peer technologies. The traditional Internet employed what is known as a client-server architecture, in which files are stored in large computers at centralized locations (servers) and end users (clients) request files from those computers. The relationship is generally regarded as hierarchical, and the amount of data uploaded by clients is very small relative to the amount of data downloaded by servers. In the classic example of the World Wide Web, client traffic consists solely of uniform resource locators (URLs), the short bits of code identifying a particular website address. Server traffic, which consists of the data comprising the requested website, is much larger. For this reason, the technologies that took the early lead in broadband deployment (cable modem service and DSL) adapted an asymmetric architecture, allocating a larger proportion of the available bandwidth to downloading than to uploading. Newer technologies, such as fiber and wireless broadband, follow the same pattern.

Peer-to-peer technologies follow a very different approach. Edge computers in a peer-to-peer architecture are not divided into those that host files and those that request files. Instead, computers simultaneously perform both functions. Because this relationship is regarded as less hierarchical than client-server relationships, the computers in this architecture are known as peers and communications between them are known as peer-to-peer. Peer-to-peer is thus not synonymous with file sharing or user-generated content, as is often mistakenly assumed. On the contrary, many peer-to-peer applications (such as Vuze) support commercial broadcast services, and many platforms for user-generated content (such as YouTube) employ centralized servers. The real significance of the term "peer-to-peer" lies in the nature of the network architecture.

It is not yet clear what proportion of network traffic will follow each architecture. For example, peer-to-peer traffic had consistently outstripped client-server traffic for several years leading up to 2007. In 2007, however, client-server traffic staged a comeback, thanks primarily to the expansion of streaming video services like YouTube, and exceeded peer-to-peer traffic 45 percent to 37 percent. Many industry observers now predict that although peer-to-peer will remain important, it will decline as a percentage of total Internet traffic over the next several years. Even so, it is clear that peer-to-peer traffic is likely to remain a more important component of network traffic than during the Internet's early years.

The differences in topology mean that traffic that is otherwise similar may travel through the network at different speeds, with different costs.

The growing importance of peer-to-peer technologies is causing significant congestion in certain areas of the network and is putting pressure on the traditional approach to pricing network services. The emergence of end users as important sources of data is putting severe pressure on the limited bandwidth allocated to upload traffic. In addition, unlike in a client-server architecture where end users usually only generate traffic when a person is seated at the keyboard, edge computers in a peer-to-peer architecture can generate traffic for as long as the computer is left running. The result is that the lion's share of upload traffic is generated by a small number of superheavy peer-to-peer users. As few as 5 percent of end users may be responsible for generating more than 50 percent of all Internet traffic.

The most recent generation of peer-to-peer technologies can exacerbate congestion still further. In the first generation of peer-to-peer technologies, each end user stored the entirety of the files that the user hosted. As a result, anyone requesting those files was limited by the total bandwidth and the level of congestion associated with the network connection attached to that end user's computer. Technologies such as BitTorrent follow a different approach. Instead of storing entire files in one location, BitTorrent divides each file into

pieces and distributes them at multiple locations around the Internet. When a BitTorrent user requests a file, the software then retrieves the various pieces from multiple computers at the same time, which reduces the amount of bandwidth required from any one peer and improves download performance. BitTorrent also dynamically reallocates requests for pieces away from the slowest connections and toward the fastest connections, thereby placing the heaviest burden on those peers with the fast connections.

The congestion caused by peer-to-peer technologies weighs heaviest on last-mile technologies that share bandwidth locally, such as cable modem and wireless broadband systems. For example, cable modem technology requires that subscribers share bandwidth with the other households operating through the same neighborhood node. As a result, cable modem customers are significantly more vulnerable to the downloading habits of their immediate neighbors than are telephone-based broadband systems, which offer dedicated local connections. Service can slow to a crawl if as few as 15 of the 500 or so users sharing the same node are using peer-to-peer applications to download files.

The classic economic solution to congestion is to set the price of incremental network usage equal to the congestion costs imposed on the network by that usage. However, determining the congestion cost imposed by any particular user at any particular time can be quite complex. Subscribers that use large amounts of bandwidth can contribute very little to network congestion if they confine their usage to hours when network usage is low. Conversely, subscribers that use only small amounts of bandwidth may nonetheless impose significant congestion costs on the network if they generate traffic at peak times. The contribution of any particular usage cannot be determined simply by counting the number of bits being transmitted. The overall impact of any particular increase in network usage can only be determined in light of other subscribers' Internet usage. Thus it may make sense to charge different amounts to users who are using the Internet to access the same content or application if a sufficient number of other users sharing the same bandwidth are using the network at the same time.

The growth of peer-to-peer technologies has also heightened the pressure on the models that network providers have used to price their services. As noted earlier, the traditional approach charges content and application providers prices that increase with the peak bandwidth consumed, while end users are charged on an unmetered basis. The fact that every download had to pass through one link that charged on a volume-sensitive basis allowed this pricing approach to serve as a reasonable approximation of efficient congestion pricing. For example, 100 downloads of a 700 megabyte movie would generate 70 gigabytes of traffic from the server, which in turn would be reflected in the price paid by the content provider to its ISP.

The situation is quite different under peer-to-peer architecture. In that case, the movie could be downloaded once from the server, and the remaining 99 downloads could be served by other end users running the same peer-to-peer software. Because end users are provided with service on an all-you-can-eat basis, the additional 99 downloads served by the peer-to-peer network do not generate any additional revenue. The only revenue received by the network is for the initial 700 megabyte download. Thus, in a peer-to-peer architecture, the amounts that content providers pay under the traditional pricing regime no longer serve as a workable approximation of the total traffic they impose on the network. Moreover, the failure to charge network participants prices that reflect their incremental contribution to congestion causes excessive consumption of network resources that ultimately harms consumers.

It thus comes as no surprise that the network providers that are most subject to local congestion are experimenting with other means for managing the congestion caused by peer-to-peer applications. For example, Time Warner has recently experimented with bandwidth caps and other forms of metered pricing. Although many network neutrality proponents have no objection to metered pricing, recent attempts to impose metered pricing and bandwidth caps have met such a hostile reaction from the network neutrality community that the network providers had to back down. That said, metered pricing is far from a panacea. As I have discussed in greater detail elsewhere, true congestion-based pricing would vary from moment to moment based on the volume of traffic introduced into the network by other users. Such a pricing regime would challenge consumers' ability to process the relevant information, and the distributed nature of the Internet means that no one entity has the information needed to formulate such policies. As a result, other network providers have turned to proxies that are strongly associated with high-volume activity, which most importantly includes a ban on operating a server as required by peer-to-peer technologies. Although this would constitute a violation of network neutrality by discriminating against a particular type of application, even network neutrality proponents acknowledge that such a restriction represents a good proxy for bandwidth-intensive activity.

Partial Transit and Paid Peering Network providers have also begun to enter into business relationships that go beyond peering and transit relationships that dominated the early Internet. Some are driven by the emergence of secondary peering relationships discussed above. Before such relationships existed, a tier-2 or tier-3 ISP would have to buy transit from a tier-1 ISP that had obtained access to all of the IP addresses that it did not serve. In other words, a tier-2 or tier-3 ISP's transit relationships would cover the entire Internet (except for its own customers).

The advent of secondary peering reduces the scope of transit services that the ISP needs to purchase. The ISP no longer needs to buy transit to the entire Internet; the secondary peering relationships already provide the ISP with the ability to reach those customers served by its secondary peering partners. As a result, these ISPs have begun to purchase partial transit that covers only those portions of the Internet not already covered by their secondary peering relationships.

In addition, an ISP with inbound traffic that far exceeds its outbound traffic may run the risk of having traffic ratios that put it in violation of its peering contract. Under these circumstances, it may attempt to cover its deficit in outbound traffic by selling a partial transit contract that covers only outbound traffic, but not inbound traffic. Alternatively, it may reduce its inbound traffic by buying partial transit for inbound traffic.

Another interesting development is the emergence of paid peering, which involves all of the same aspects as conventional peering relationships. Peers announce to the rest of the Internet the addresses that their peering partners control, maintain a sufficient number of interconnection points across the country, and maintain the requisite total volume and traffic ratios. The key difference is that one peering partner pays the other partner for its services.

Paid peering is driven by both supply-side and demand-side considerations. Starting first with the supply side, settlement-free peering arrangements between tier-1 ISPs with similar traffic volumes make sense only if both networks have similar costs. Over time, backbones have begun to serve two different types of last-mile networks: those such as Cogent and Abovenet that primarily serve content and application providers (which are sometimes called "content networks"), and those such as Comcast and Verizon that serve end users (which are sometimes called "eyeball networks"). The costs of the first type of network are quite low, typically only requiring a single high-speed line to a small number of business locations. The costs of the second type of network are considerably higher, requiring the wiring and upgrading of equipment in entire neighborhoods. The presence of such asymmetric costs provides a substantial impetus for cash to flow from networks serving content and application providers to networks providing connections to end users.

These supply-side considerations are reinforced by demand-side considerations associated with the economics of two-sided markets, which illustrates the potential benefits of allowing network providers to charge differential prices to both end users and content and application providers. Conventional economics has long recognized the existence of "network economic effects," which cause a network to increase in value as the number of users connected to it increases. To use a classic example, the value of a telephone network to a particular consumer depends in part on the number of other subscribers connected to the network; the more people you can reach through the network, the more valuable it becomes.

The benefits created by the network economic effect for telephone networks arise with respect to a single class of customers. When a market is two-sided, instead of bringing together a single class of similarly situated users, networks bring together two completely different classes of users. In those cases, the value is determined not by the number of users of the same class, but rather by the number of users of the other class. A classic example is broadcast television, which brings together two groups: viewers and advertisers. Advertisers gain no benefit (and if anything suffer a detriment) from belonging to a network with a large number of other advertisers. The value of the network for advertisers is instead determined solely by the number of viewers, i.e., the size of the other class of users.

The literature suggests that social welfare would be maximized if the network provider were permitted to price discriminate on both sides of the two-sided market. It also suggests that the prices paid on each side of the market can differ widely, and that in many cases it is economically beneficial for one side to subsidize the other side. The fact that the Internet has become increasingly dominated by advertising revenue paid to content and application providers suggests that it may be socially beneficial for content and application providers to subsidize the prices paid by end users. An advertiser's willingness to pay for an ad on a particular website depends on the number of end users viewing that website. Under these circumstances, the optimal solution may be for the website owner to subsidize the total number of end users by making payments to the network provider to help defray their costs of connection. The costs of subsidizing more users would be more than offset by the additional revenue generated by the fact that advertisers can now reach more potential customers. In the case of broadband, this would be both economically efficient and would be a boon to consumers both in terms of providing service in more geographic areas and in reducing the prices that consumers pay.

These dynamics are again well illustrated by broadcast television. In many ways, broadcast television and the Internet are analogous. The studios that create television programs play a similar role to content and application providers. Television networks aggregate programs and deliver them nationally in much the same manner as content networks and backbone providers. Local broadcast stations provide last-mile connectivity that is quite similar to the role played by eyeball networks. In addition, the revenue structure is quite comparable, in that television networks receive advertising revenue

Social welfare would be maximized if the network provider could price discriminate on both sides of the two-sided market.

in much the same manner as content and application providers. Furthermore, the cost structure is somewhat similar in that connecting individual homes is much more costly than distributing programming nationally.

For decades, the standard business arrangement has been for television networks to subsidize the operations of local broadcast stations by paying them to be members of their television networks. The industry's revenue and cost structure make such arrangements quite logical. The cost of paying these broadcast stations to affiliate with a network is more than offset by the increase in advertising revenue made possible by the fact that the network is now able to reach a larger audience. Broadcast television thus represents a prime example of when firms operating on one side of the market find it economically beneficial to subsidize end users on the other side of the market.

Furthermore, the magnitude of the affiliation fees that the networks pay to broadcast stations is anything but uniform. The precise amount varies with the relative strength of the network and the relative strength of the broadcast station. Stronger broadcast stations receive more, while weaker ones receive less. Equally interesting is the fact that in recent years, the cash flow has begun to vary in its direction as well as magnitude, with weaker stations having to pay rather than being paid to be part of the television network. The dynamic nature of this pricing regime benefits consumers by providing incentives for networks to invest in better quality programming and by providing an incentive for stations to provide better carriage.

The two-sided market analysis reveals the potential drawbacks of preventing network providers from charging differential prices. As a general matter, pricing flexibility makes it easier for network providers to recover the costs of building additional bandwidth. Granting network providers pricing flexibility with respect to content and application providers should reduce the percentage of the network costs borne by consumers. Conversely, preventing network providers from exercising pricing flexibility with respect to content and application providers would simply increase the proportion of the network costs that providers must recover directly from end users. This simultaneously raises the prices paid by consumers and decreases the likelihood that the capital improvements will ever be built. Charging content and application providers differential prices thus has the potential to increase social welfare and can reduce, not increase, the burden borne by consumers.

CONCLUSION

It is all too easy to forget that the Internet is not a monolith with a brooding omnipresence overseeing the entire system. Instead, it is a collection of autonomous systems that determine the terms of interconnection between them through a series of arms-length negotiations. Given the Internet's essence as a network of networks, it should come as no surprise that no two packets will pay the same amount for the same service.

The developments that I have outlined in this article have made such differences even more likely. The network no longer adheres to the rigid and uniform hierarchy that characterized the early Internet and its predecessor, NSFNET. Data packets can now travel along radically different paths based on the topology of the portion of the network through which they travel. This is the inevitable result of reducing costs and experimenting with new structures. At the same time that network providers are experimenting with new topologies, they are also experimenting with new business relationships. Gone are the days when networks interconnected through peering and transit and imposed all-you-can-eat pricing on all end users. That fairly simple and uniform set of contractual arrangements has been replaced by a much more complex set of business relationships that reflect creative solutions to an increasingly complex set of economic problems. Again, these differences mean that the service that any particular packet receives and the amount that it pays will vary with the business relationships between the networks through which it travels. Although many observers reflexively view such deviations from the status quo with suspicion, in many (if not most) cases, they represent nothing more than the natural evolution of a network trying to respond to an ever-growing diversity of customer demands. Imposing regulation that would thwart such developments threatens to increase costs and discourage investment in ways that ultimately work to the detriment of the consumers that such regulation is ostensibly designed to protect. R

Readings

- "Advances in Routing Technologies and Internet Peering Agreements," by Stanley Besen et al. *American Economics Review*, Vol. 91 (Papers and Proceedings) (2001).
- "Assessing the Impact of BitTorrent on DOCSIS Networks," by James J. Martin and James M. Westall. *Proceedings of the 4th International Conference on Broadband Communications, Networks and Systems* (BroadNets), 2007.
- "Bringing Neutrality to Network Neutrality," by Kai Zhu. *Berkeley Technology Law Journal*, Vol. 22 (2007).
- "Network Neutrality after *Comcast*: Toward a Case-by-Case Approach to Reasonable Network Management," by Christopher S. Yoo. In *New Directions in Communications Policy*, edited by Randolph J. May (Carolina Academic Press, 2009).
- "Network Neutrality and the Economics of Congestion," by Christopher S. Yoo. *Georgetown Law Journal*, Vol. 94 (2006).
- "Overlay Networks and the Future of the Internet," by Dave Clark et al. *Communications and Strategies*, Vol. 63 (2006).
- "The Flattening Internet Topology: Natural Evolution, Unsightly Barnacles, or Contrived Collapse?" by Phillippa Gill et al. In *Proceedings of the 11th International Conference on Passive and Active Management*, edited by Mark Claypool and Steve Uhlig (Springer, 2008).
- "The Growing Complexity of Internet Interconnection," by Peyman Faratin et al. *Communications and Strategies*, Vol. 72 (2008).
- "Was the Internet Ever Neutral?" by Craig McTaggart. Paper presented at the 34th Telecommunications Policy Research Conference, 2006.