

**Shaping Infrastructure and Innovation on the Internet:  
The End-to-End Network that Isn't**

Christian Sandvig (csandvig@uiuc.edu)  
University of Illinois at Urbana-Champaign

Author's draft as submitted for publication as:

Sandvig, C. (2006). Shaping Infrastructure and Innovation on the Internet. From: D. H. Guston & D. Sarewitz (eds.), *Shaping Science and Technology Policy: The Next Generation of Research*, pp. 234-255. Madison, Wisc.: University of Wisconsin Press.

\*Winner, "The Next Generation of Leaders in Science and Technology Policy,"  
A junior faculty competition funded by the US National Science Foundation and co-sponsored by Rutgers University; the Center for Science, Policy, and Outcomes (CSPO); and the American Association for the Advancement of Science (AAAS).

## **Shaping Infrastructure and Innovation on the Internet: The End-to-End Network that Isn't**

This chapter approaches the question of how we should best reason about the design of communication infrastructures by examining a particular debate about the Internet: specifically at issue are the benefits of the Internet for innovation. Some argue that the Internet's gift is found in an obscure design feature called the "end-to-end argument."<sup>1</sup> This is a network engineering strategy that promotes "stupid" networks: designs where the center lacks intelligence and performs only a few functions, while nodes at the edge of the network—the *ends*—build complex applications by employing the simple building blocks of the core. This is the Internet (smarter PC, dumber router) as the opposite of the telephone (dumber telephone, smarter switch). Proponents say that with the Internet's end-to-end design, experiments can be deployed from the edges (ends) by anyone at all. Success of the Internet can be explained because experiments like the World Wide Web did just that.

On the other side are commercial interests currently deploying intelligence inside the network's core; this logic speeds some traffic over others (caching), blocks traffic (firewalling, filtering), eavesdrops (snooping), and disguises some nodes as others (masquerading).<sup>2</sup> This is so worrying for innovation—and for the freedom of users—that end-to-end proponents have asked the US government to take action to preserve the Internet's "natural" form. They argue that we may even need a new Internet that retains an end-to-end design if the present one continues to erode.<sup>3</sup>

This chapter takes a third position. Forward from the earliest organized communication systems such as the horse couriers of the early Chou dynasty, circa 1000 B.C.E., history teaches that networks tend to complexity at the core as more is asked of them. Contrary to the end-to-end argument, there is no reason to think that the Internet will evolve to be faster and more reliable than earlier electronic systems, and in a reversal of three millennia this will require fewer intermediaries and less intelligence at the center.<sup>4</sup> While a more complex “middle” is already here, the question at issue has never been the preservation of a simpler network structure but how and where new complexity is implemented. The key to innovation rests somewhere else entirely: the key is not the degree of logic within intermediary nodes, but which nodes we trust. While Moors (2002) elaborates some of the technical implications of this critique, this chapter addresses the implications for innovation and public policy.

I will address the future Internet by first recalling the oldest data networks – chiefly to remind us that though the structure of a communication network may have a technical veneer, it is a political bargain. Then, considering the Internet, I will unpack the end-to-end argument and suggest that: (a) it is not an organizing principle, (b) if it is a principle it is probably not true, and (c) if it is true it is probably not useful. The best outcome that normative claims premised on the end-to-end argument can offer us is to produce the right result for the wrong reasons, but we might be even better at promoting innovation if we act for the right reasons. Even worse, a dogmatic belief in end-to-end will simply retard the development of the Internet’s infrastructure by limiting needed improvements in the “middle” or core. I will suggest that the right values to support are transparency and participation. I then conclude with the suggestion that underlying

principles that support innovation need to be addressed explicitly, not silently embedded in technical arguments.

### Internet Design is an Old Problem

To tame the policy case of the present Internet, let us recast the present problems of high technology in the relatively sedate terms of technologies long dead: imperfect, but at least relatively settled examples. Data networks have a long history – the first things we might qualify with the term “systems” of communication were made from human couriers. The first courier networks were point-to-point, with most of the intelligence about the network’s condition located at the network’s edge – there were no “courier network commissions” or planners of overall “infrastructure.” Messengers relied on their own knowledge of existing paths. Just as the early Internet was a somewhat unreliable component of a computer-based communication system, once an ancient messenger left for his destination it was by no means assured that he would arrive. There were often financial and political benefits to intercepting messages, and “courier loss”<sup>5</sup> could also be a simple case of highway robbery. As couriers are much harder and slower to replace than Internet packets, the solution was not to re-send lost messengers, but to change the network itself.

This complexity then distinguishes between two generations of ancient courier systems. The first generation system involved only couriers using existing paths: whatever route was at hand. Second-generation courier systems had a greatly improved capability for signaling the state of the network, improved security, better reliability, and better performance. The first

second-generation courier system known arose in China early in the Chou dynasty (approximately 1000 B.C.E.). Typically, rather than relying on the paths that already existed, second-generation networks incorporated “sponsored” roads or “post” roads that were paid for through taxation. Money spent on marking and improving the quality of the road allowed the traffic to travel faster – examples abound from the Roman *viae militares* (the military road) to the Spanish *El Camino Real* (the king’s highway). But often these were not just upgrades to the same old infrastructure – beyond improvements to the old routes, the second-generation courier network added intelligence that was not at the network’s edge.<sup>6</sup>

The Chinese system introduced repeaters: a system of “post-houses” allowed tired couriers to pass their messages to fresh, rested ones. In Chinese, Persian, Roman and other civilizations, runner relays were later supplanted by couriers on horseback. Not only did the post-house or relay system increase the speed messages could travel, post-houses were also used as a place to post guards to protect the integrity of the network from spies and robbers. They served a routing function by directing riders over alternative paths and they maintained information on the quality of the routes. In some incarnations they billed for service, as toll stations. They filtered traffic – some post-house systems included armed guards and a capacity to inspect messengers. While Babylonian Royal Couriers were allowed to pass, the Bedouin raiders that hunted them most certainly were not.<sup>7</sup>

One of the most ingenious additions to second-generation courier systems was the fire beacon. If the guard posts were numerous and located within sight distance, in clear weather the fire beacon could be used to send a prearranged message much faster than even a mounted

courier, providing the network two modes of operation. These beacons could signal meta-information: in some systems they were used by the post-houses as a “trouble” signal warning of a problem with the network itself (a signal from the network’s center to its edge, or as coordinating signal between different post-houses in the center). In special circumstances this faster mode could be quickly repurposed to carry simple information – often, warning of invading armies. The integration of the beacon into courier networks (a proto- semaphore) marks not just a general increase in the sophistication of the system but also the advent of segregating traffic into classes by priority, where each class has a different form and a different quality of service.<sup>8</sup>

Policy implications of second-generation courier networks. As it entailed investment in roads, the second-generation courier network required much greater standardization of traffic. Wheeled vehicles were nothing new – they predate the Chou second-generation courier network by about 1000 years – but when roads were improved the road width had to be decided and future traffic had to conform. This was a decidedly local practice: at least 84 inches in width was required for a Roman road, but only 55 inches between shaped-stone wheel ruts sufficed for Greek sacred roads. These standards were then imposed on users by technology (the road), but also by law: the *raeda*, the fast freight cart used on Roman roads, was restricted by law to carrying 750 lbs. or less, for fear of damaging the road.<sup>9</sup>

Second-generation networks must have been much more effective than their predecessors and many orders of magnitude more expensive: guards, post houses, and roads all cost money, but there were fewer dead messengers. The trend in the evolution of ancient courier networks is

clear. As the demands on the networks increased, they became faster and more reliable in part by becoming more elaborate. As rulers added resources to the infrastructure, the communication network designers of the ancient world also added intelligence to the center of the network to manage and control it. They asserted this increasing control through a combination of force, law and technology: the network's soldiers, rules and form. As courier networks improved, the number of intermediaries must have increased at least slightly: no longer were couriers left on their own to wander, in the improved systems the message might be normally expected to pass through many hands to be relayed, taxed, and inspected.

The Internet and the unexpected reversal of history. In a hotly contested debate now occurring somewhere between network design and public policy, some argue that the optimal evolution of electronic networks will be just the opposite of these ancient courier networks – the Internet will evolve to be faster and more reliable than earlier electronic systems, and this can, in a reversal that is both paradoxical and exciting, require fewer intermediaries and less intelligence at the center of the network. In a sense, much of the early excitement about the Internet stems from this controversial design concept; it is the spring from which ideas about the Internet's “inherently decentralized nature” flow. In the remainder of this chapter I will unpack this idea, known as the “end-to-end argument,” and then reassess it, employing the examples given above.<sup>10</sup> As additional examples will show, end-to-end has increasingly moved from engineering circles to public policy circles. The focus of our interest here is the advocate's perception that the end-to-end argument in network design, usually evaluated by engineers on the basis of technical efficacy, may also be normatively positive: a kind of philosophy of functional

network topography. Normative benefits have been claimed for two reasons: First, under some conditions end-to-end networks can dramatically increase the number and diversity of groups that can participate in the design of network applications. Second, end-to-end networks can make it more difficult for unwanted third parties to control communication on the network. In this way, the end-to-end network has been hailed as inherently more democratic, its form producing user freedom.

However, I will argue that the current technical and political debates about end-to-end are misleading, for they cloak arguments about power with appeals to a single objective technical truth where none exists. Indeed, it is not end-to-end design per se that is normatively positive, but the transparency, openness, and participatory design consultation that have come to be associated with this model of network intelligence through history and tradition.<sup>11</sup> Loading the end-to-end argument with these social goals (rather than addressing normative goals directly) is a dangerous and misguided strategy because it shifts policy discourse away from normative ends in favor of traditional means that may no longer lead where we expect.

What is an End-to-End Network?

As is common in many areas of technological development, computer system and computer network development since its inception has exhibited a trend toward modularization. What was once a single piece of technology (“the computer”) becomes an assemblage of different parts, some of which are standardized and produced by different parties. The innovations of the Internet reside in software, but modularization has still proceeded in a manner

similar to the standardization of parts for a car. Today's programmers don't start a project writing binary codes to control the hardware of a computer, instead they assemble software by relying on a variety of standard components such as code libraries and features of the operating system. The modularization of software has been tremendously positive: it facilitates innovation by allowing system developers to incorporate standardized components in new and larger projects, it increases reliability in software by distributing what is effectively pre-packaged expertise, it reduces software development time by changing software development from a process of construction to a process of assembly, and it tends to drive down software cost by allowing competition in development and provision of these modular subcomponents. The clearest expression of modularization in computer networking has been the development of a standard model—called the seven-layer model—for representing communication between computers.<sup>12</sup> This model effectively divides the task of communicating into sub-tasks (layers), making the development of new network applications such as Web browsers and e-mail clients much simpler because the application programmer can leave most of the job of communicating to other subsystems. The entire means for sending messages does not require reinvention each time a new kind of message needs to be sent.

Elaborating this last example of modularization, the seven-layer model, we see that it is useful for reasons of efficiency because it saves development cost. But as networking and the idea of network layers became widespread in the 1980s, so too did disputes about the technically correct level within a network to locate certain functionality: “should this software company be responsible for a given task, or should that one?” In a now-classic paper in network engineering,

Saltzer, Reed, and Clark elaborated a design strategy for computer systems called the “end-to-end argument” (1984).<sup>13</sup> End-to-end arguments are those that take the position that:

- (i) If a particular function requires the participation of the endpoints of the system, it should not be implemented in any other location in the system.

This is the first of six principles or statements that I extract from arguments made on behalf of end-to-end networks. This statement can be explained for the functional goal of *reliability* with a somewhat oversimplified scenario of mailing postcards.<sup>14</sup>

If person A mails a very important postcard to person B, how does person A know that the postcard has arrived safely? One option would be to apply a variety of strategies at points in the postal system where loss of the postcard is likely to occur: the location of each item of mail in the system could be constantly tracked, duplicate postcards could be made at several points and delivered as insurance, postcards could be fabricated out of a durable metal instead of flimsy paper. The end-to-end argument, however, notes that while these strategies would increase the reliability of the postal system as a whole, they are expensive and might tend to slow the delivery of all mail. Indeed, regardless of any of these costly improvements to the postal system, it could never be canonically assured by person A that person B did in fact receive a particular postcard. To attain peace of mind, person A would still have to ask person B if the postcard had arrived. It is likely more efficient, then, to skip the improvements to the postal system as a whole and for person A should plan on asking person B. As long as the use of the post is cheap and simple, if

some of the postcards do not arrive, person A can ask for them to be sent again – by, for instance, using another postcard. In this scenario, person A and B are the “ends” and the intelligence necessary to confirm receipt of the message resides entirely in the ends of the network. No complexity was added to the middle (or core) of the postal network, instead it was the behavior of the “devices” connected to the network—the users—that changed to support the goal of reliability. In the eyes of the end-to-end proponents, simpler networks and smarter ends are always the answer.

A competing strategy: The end-system model. The early telephone network, in contrast, was initially developed with a network design philosophy that has been called “end-system” (Kruse, Yurcik, and Lessig 2001).<sup>15</sup> Most of the functionality of the network was located in the telephone switch, and the “end” (the telephone) was a fairly simple device that could do little except relay simple commands to the switch (off-hook, on-hook, 1234567890\*#). To expand slightly on the postcard example above, we see that the end-to-end argument is clearly about who does what. If the communication problem is to insure that a postcard sent between person (A) and person (B) arrives, some end-to-end strategies might be that person (A) keeps very careful track of the postcard, (A) makes a duplicate of the postcard in case it is lost, or (A) chooses to make the important postcard out of a very durable material. In comparison, the end-system strategies would be that the post office implements a tracking system to keep very careful track of all postcards, the post office makes duplicates of all postcards when they are mailed, and the post office requires that all postcards be made more durable. And so in computer system design

these questions about the location of functionality can seem like purely technical concerns, but the questions drawn from this example are also clearly “who is in control?” and “who pays?”

End-to-end arguments have waxed and waned. When end-to-end was first proposed by Internet pioneers it was controversial – end-system designs ruled the day in computer communication. In the related sphere of telecommunications, the dominant thinking was later called the “intelligent network” movement (Mansell 1994). In the late 1990s, backlash against this centralization led to a rediscovery (or independent discovery) of end-to-end in telecommunications under the catchphrase “the dumb network” or “the stupid network” (Isenberg 1997). Note that the categories of “end-to-end” and “end-system” are useful, but they are not exhaustive, they are not a dichotomy, and indeed may not even be on a continuum. It is possible to characterize the structure of some networks as predominantly end-to-end or end-system, but other networks defy this characterization – they are confusing hybrids.<sup>16</sup>

Nevertheless, and despite the danger of over-generalization, table one attempts a rough categorization by way of example. As “end-to-end” and “end-system” networks are design strategies and ideal types, no network can clearly be labeled one or the other, but in this table I attempt to match the ideal types with a few examples that may come closest.

[Insert Table 1 here]

The earliest ad-hoc courier networks were end-to-end in that the network was so simple that no complexity existed, except at the edge. All applications (diplomacy, commerce) were

bargains between end-points to which the center of the network did not contribute. Although the earliest networks of paths were probably not, strictly speaking, “designed,” still no intermediaries at all were planned. (Bandits are an example of an unplanned intermediary.) The earliest, non-switched “party-line” telephone systems could also be conceptualized as end-to-end. Really they are broadcast networks, as all terminals received all transmissions, and there was no complexity at all in the middle of the network. This brings us to a point of some confusion: the implied relation between end-to-end and switching. Note that the distinction between end-to-end and end-system networks is not related to switching. Switching is just one function of a network that might be implemented with an end-to-end or end-system strategy. For instance, the Internet is a packet-switched network that end-to-end proponents worry is becoming less end-to-end—but it will still be packet-switched. A broadcast (non-switched) network could implement many functions in a complicated network core that are unrelated to switching. Last in our table, we have the network that gave birth to the concept of end-to-end: the recent Internet before the advent of technologies that allegedly “break” end-to-end (discussed later).

In the right column, the Chinese “post-house” system of the Chou dynasty could be considered end-system if we take some liberties with the analogy between courier systems and data networks. The post-house system also implemented routing at the core, but in addition it probably supported spying, taxation, and some other functions. It is a difficult example for this ideal type.<sup>17</sup> We also have the network where the concept “end-system” originated, plain old telephone service before the advent of technologies that increase intelligence at the terminal

device, or end, and muddle the example (such as ISDN). While the Internet's core handles routing (an example of some intelligence at the center), the Internet is still considered an end-to-end network because little other function was implemented in routers at the core.

This broad distinction between two ideal types in overall strategy for communication network design has consequences far removed from the practice of computing. The design of any technology that allows humans to communicate must have social and political consequences, but more specifically the perceived benefits of end-to-end design have been put forward in the case of the Internet in three areas: (1) user-driven innovation, (2) protection from unwanted intermediaries, and (3) technical correctness. I will now discuss each in turn.

The implications of end-to-end for user-driven innovation. Advocates of the end-to-end approach point out that it reduces the complexity of the "core" network, and that the resulting generality of the network fosters innovation. That is, it is more likely that some new and unanticipated service can employ the basic building blocks of a simpler network core (Blumenthal and Clark 2001). The innovation principle could be embodied succinctly as:

- (ii) The lowest layers of a system should provide the greatest flexibility possible, so as to permit applications that cannot be anticipated.

The Internet is a manifestation of this principle: it provides a fairly general set of facilities to allow the transfer of "data." A variety of different applications have emerged (e.g., remote login, electronic mail, the World Wide Web) that all share the same Internet. If certain

functionality is required to allow one of these applications to work, it is located in software (the electronic mail client, the Web browser) on computers attached to the network's "edge" (Blumenthal and Clark 2001, 92).

It is a key feature of end-to-end design that *users* of the network are able to create new applications – applications that eventually drive the technology itself (cf. von Hippel 1988; Bar et al. 2000; Bar and Riis 2000).<sup>18</sup> This process of "user-driven innovation" occurs with other technologies, but the fact that the control mechanisms for the Internet reside in software allows user-driven innovation a much more central role. Consider that electronic mail and the World Wide Web (the two most popular Internet applications at the time of writing) were developed not by a centralized network authority or the owner of the network, but by users: e-mail by Ray Tomilson for his group at BBN in 1973, the Web by Tim Berners-Lee for his group at CERN in 1990.

By contrast, in an end-system model the device at the edge of the network (e.g., the telephone) is simple and inexpensive to manufacture. The network equipment to which the device connects provides all of the intelligence and functionality of the system. In the early telephone network, adding new technology at the end was expressly forbidden—adding an answering machine or even a piece of cardboard was a violation of the system's design principles, and possibly illegal.<sup>19</sup> Both the design philosophy and the difficulty in configuring hardware act against user-driven innovation on the telephone network. If a user of the telephone network wished to implement a new service such as three-way calling, before electronic switching this would require finding a screwdriver and linking three wires. Even after electronic

switching and software control, if a user wished to implement a new service on the telephone network such as voicemail, no facility exists to allow it: the system is controlled by the telephone companies and the user may not tamper with it—or even discover how it works—without permission.

End-to-end as protection from intermediaries. Those with technological determinist leanings have pointed out that the structure of the end-to-end Internet is itself a protection from anyone who would limit the freedom of communication. By this logic, adopting an end-to-end Internet design will produce freedom irrespective of legal or social arrangements, or substituting for them. For much the same reasons that would-be innovators can deploy any new application that suits their fancy, would-be communicators do not need permission to speak and need not subject their speech to anyone’s review. This argument says that if the network has no functionality to examine the messages it carries, communication is then more free.

The argument from technical correctness. Some have sought to put forward end-to-end design principles as true in some objective sense. Principle (ii) is less overtly concerned with objective correctness, but the next arguments we will consider are related to the idea that if functionality is removed from the network core, the core itself becomes easier to administer, likely cheaper, and faster. They argue that a simpler core is necessarily more transparent and easier to model (for an account and critique of this specific point, see Moors 2002). These “correctness” arguments have been made and re-made in the landmark papers on end-to-end, in fora like the end-to-end mailing list, and likely in engineering meetings around the globe with

increasing frequency over the last twenty years. It has been claimed that end-to-end is technically correct because:

- (iii) Any function implemented in the core network may be redundant because this functionality has already been implemented at the end-point.
- (iv) Any function implemented in the core network may be redundant because some applications will never need it.

Let me return again to the postcard example to explain these claims of end-to-end proponents. To illustrate principle (iii), consider that if *both* the end-to-end and the end-system strategies in table one are implemented, there will be a significant duplication of effort. Furthermore, to illustrate point (iv), imagine person C who does not need to verify if his postcards are delivered or not (or does not wish to pay extra for it). If person C pays for the postal system (through stamps or taxes), part of his contribution to the postal system through taxes will go to support improvements he does not need or want. The network will be more expensive to operate, with no benefit to him, say the end-to-end proponents. A stronger form of statement four is that *no* function should be implemented in the network unless *all* clients of the network (or that network layer) will need it, because:

- (v) The end-point tends to have more information about what it needs than the network.

- (vi) Any function implemented in the core network adds cost and complexity that is borne by all network users, even if they do not use the function.

Principle (v) clearly implies a particular imagined user and a particular network. While it seems to be an argument for end-to-end design, it presupposes an end-to-end network where there is no central authority dictating what applications should be used, and where the technology and form of communication remains unsettled. If the Internet will continue to be a place where new applications arise continually (e.g., as peer-to-peer file sharing in the form of Napster recently arose) there is some merit to this point. Statement (vi) is merely an articulation of the conclusions arrived at earlier with the postcard example.

#### The Perceived Challenge to End-to-End

Attempts to promote new Internet applications in the last eight years have proponents of end-to-end design feeling as though they are under siege. Proponents have warned that developments both in software, policy, and use are “compromising the Internet’s original design principles” (Blumenthal and Clark 2001). Blumenthal and Clark note four examples of emerging requirements for Internet applications that challenge end-to-end design principles: (1) the need to manage untrustworthy end-points, (2) demands for better throughput required by streaming audio and video, (3) differentiation of service between competing Internet Service Providers (ISPs), and (4) the rise of increasing third-party involvement in communication. That is, each of these might imply a need to add intelligence to the network: for example, (1) the “firewall” to block

“hostile” network traffic, (2) the addition of proprietary distributed cache systems for multimedia content,<sup>20</sup> (3) distinguishing between different kinds of content by the ISP to provide different levels of service quality, and (4) the use of filters to block unwanted (or illegal) traffic - or traffic analyzers to eavesdrop on suspect traffic.

The perceived danger of these changes is that they each constitute new intermediary points in the path of traffic traversing the network where some third party may exert control. As mentioned above, the early enthusiasm for the Internet as a communication medium could be seen as an excitement about the lack of intelligence inside the network. Claims that the Internet is “fundamentally” anti-hierarchical, free, democratic, decentralized, or impossible to regulate fall within this scope. Those concerned about the censorship of content or the leveraging of the control of the Internet’s wires into the control of its content have pointed out that “end-to-end was initially chosen as a technical principle. But it didn’t take long before another aspect of end-to-end became obvious: It enforced a kind of competitive neutrality. The network did not discriminate against new applications or content because it was incapable of doing so” (Lessig 2000).

Indeed, one of the original authors of the seminal end-to-end paper equates end-to-end with “the default situation [where] a new service among willing endpoints does not require permission for deployment.” Abrogation of end-to-end design principles leads to the case where “new chokepoints are being deployed so that anything new not explicitly permitted in advance is systematically blocked.” (Reed 2000, 4).

The strength of the arguments made to ensure future end-to-end design rely on the presumption that (1) end-to-end *is* the current design scheme, and that (2) the current design scheme has been an effective one. Therefore, don't fix it if it isn't broken. As Lemley and Lessig state, "We do not yet know enough about the relationship between these architectural principles and the innovation of the Internet. But we should know enough to be skeptical of changes in its design. The strong presumption should be in favor of preserving the architectural features that have produced this extraordinary innovation." (Lemley and Lessig 2000, 4).

These appeals have been made in policy fora, and seek to constrain those who would depart from end-to-end design principles. These threatening parties are private actors (such as ISPs and telecommunications companies) free to attach their new software and equipment to the Internet as they see fit. In other words, end-to-end proponents wish to make the case that while the Internet has a "fundamentally" decentralized and distributed nature, still it now requires policy action to prevent private actors from departing from its technical design principles.

The most visible recent manifestation of this conflict has been the US cable industry "open access" debate. Cable providers in the US have deployed extensive intelligence in the network's "middle": they have deployed a technology called a caching gateway at the junction of their subscriber network and the slower Internet. As cable providers already own governmentally-sanctioned monopoly franchises, if they are allowed to require that users of broadband cable modem service also use an ISP that they own, it is likely that they will be tempted to leverage their monopoly power into control of Internet content. This could be done by hosting "strategic partners" or in-house content on these caching gateways that they alone

control and that are close to cable modem subscribers. This would provide subscribers with faster access to content that generates profits for the cable company, and indeed also provides the cable company with a temptation to slow traffic from competing entities. The US Department of Justice, the Federal Trade Commission, and the Federal Communications Commission were lobbied with this rationale to require open access to competing ISPs in merger proceedings involving large cable providers. For a review of this example, see Bar, Cohen, Cowhey, DeLong, Kleeman, and Zysman (2000).

End-to-end: already over or never was? In this section I hope to further critique end-to-end approaches in order to show that the end-to-end debate is not one of technical correctness, but one of sociopolitical control. First, I should point out that what has been presented above is a classic case of a technology in the early period of its development, when its inner workings are subject to interpretive flexibility (Pinch and Bijker 1987). Several relevant social groups have been identified in this controversy, e.g.: private firms that seek to make a profit from the Internet in some way and the “old school” of Internet pioneers, designers, and computer scientists. The Internet is a technology that they each seek to shape, and to assert control over it they are engaging in a definitional controversy about what features are essential to the object “Internet.” They attempt to control the emerging technology by painting the encroachments of other groups as antithetical to the natural form of the thing that we call “Internet.”

As the authors of the original end-to-end paper note, end-to-end is an “argument,” rather than a rule, law, or even principle. Determining what is meant by “ends” in any given engineering problem is extremely difficult and the resulting debates are open to much

interpretation. Even the strongest case made for end-to-end, the case for the promotion of innovation, is problematic. As has been pointed out, the end-to-end approach only facilitates innovation of the kind where the new application services can be built with the low-level building blocks provided by the network. As network engineers do not agree on a kind of “periodic table” of low-level network building blocks with which every possible service can be built, the end-to-end goal of a simple and easy-to-build-from network is actually an easy-to-build-from network where it is easy-to-build only certain kinds of things. It is easy to build things that are made from the blocks or functions that are present. If the service you want to build cannot be built from these functions, an end-to-end design strategy will not make your new idea easier to build.

For instance, those who wish to introduce robust multimedia streaming services object that it is not possible to build quality of service guarantees when starting from the building blocks that are available. To work at all, the development of streaming for multimedia content seems to require changes to the core functionality of the network. It is hopelessly impractical to treat broadcasting as a point-to-point operation in the present Internet. With any large number of users wanting to view a broadcast stream simultaneously, the load on the provider of the stream quickly becomes unmanageable, and the network near the provider is subject to congestion. This is because the provider must send multiple identical copies of the stream, one for each user that desires it – an ugly approach and an inefficient one. The stupid network does not notice that each stream is the same thing and cheerfully sends a thousand copies when one would be best.

The cooperative open approach to redesigning the Internet's core to support this semi-broadcasting of multimedia content is termed "multicasting," and it involves modifying the Internet's basic protocols to reduce the duplicate transmission of multiple streams where one stream would do. These core network modifications have been described by end-to-end proponents as a necessary departure from an end-to-end strategy for reasons of performance.

However, recently providers of private distributed stream-caching services have begun to offer reliable broadcasting on the Internet without modification of existing protocols or the network's core. These providers, such as Akamai and formerly Inktomi, locate servers in many data centers around the world. The content stream is sent first to these data centers, where it is then duplicated for users that are nearby. In other words, users are directed to obtain the content from the point that is nearest them. This approach has been decried by end-to-end proponents as a violation of end-to-end.

Clearly the approach used by Akamai reduces the transparency of the network. Akamai is a private company and it has introduced a proprietary facility for providing content that is available only to its subscribers. In contrast, if multicasting were fully realized within the Internet's protocols, this would provide a facility inside the network for providing such content that could be used by anyone. In effect, a key difference between these two approaches is cost distribution. Modifying the core network, as Akamai does not do, distributes the cost of broadcasting Internet content across all users: those who want the content, those who provide it, and those who never use it. The Akamai approach requires that multimedia content providers and providers of popular (non-multimedia) content pay for this infrastructure as an added service, in

advance. The decision of where to locate the functionality for broadcasting determines how Internet broadcasting is funded and who has access to it.

### The Problem of Control

In considering the design of the Internet, it is easy to despair when a decision about network design must be made without much precedent. But the history of communication networks abounds with useful comparison problems in the distribution of intelligence within the network. Indeed, the present debate can be usefully informed by the history of courier networks and the historical difficulties of asserting control through network protocols. The word “Internet” evolved from what ARPANET protocol designers called “the Inter-network problem.” That is, the between-network problem: “Networks represent administrative boundaries of control, and it was an ambition of [the ARPANET] to come to grips with the problem of integrating a number of separately administrated entities into a common utility.” (Clark 1988, 107). The second-generation courier network of the Chou dynasty that I described above, then, needs to be extended further still – an apt comparison is rather the dilemma of what happens when the couriers reach the edge of a kingdom and must pass onto courier networks controlled by other kings. To borrow the language of networking, the network designers of ancient China did not consider technical means for ensuring the integrity of their communication across these administrative boundaries of control, because any protocol they might devise would depend entirely upon the cooperation of rulers of neighboring kingdoms. If it was in the strategic interests of other rulers to intercept or otherwise compromise messengers, the problem is not a

technical one but a one of politics. The only way to ensure that adjacent interconnected networks behave the way you want them to is then to assert control over them in some fashion.

The role of “technical” argument. The use of end-to-end as a design principle has the effect of pushing intelligence to the borders of the network. This creates what end-to-end advocates argue are “application insensitive” networks; systems that are optimized to deliver few “low-level” services at the core—it has been claimed that these networks offer building blocks upon which many types of use can be built, and an environment where there need be few intermediaries. I hope to have demonstrated that this “low-level” of service is better described as one form of service among many possibilities: the Internet is application insensitive as long as the application is similar to what was envisioned by the first designers of TCP/IP. History teaches us that the Internet has always had intermediaries, but that end-to-end proponents were happiest with the intermediaries that they knew (e.g., service providers – initially universities). The end-to-end argument is a way to stop the new intermediaries by arguing that they are technically incorrect – a definitional debate about the form of the technological artifact “Internet” which has not yet stabilized.

I do not mean to underestimate engineers: the above does not mean to suggest that those who make end-to-end arguments as technical solutions are in any way simple. Many of the participants in these debates are well aware of the dangers of advancing normative principles as technical principles, or conflating the technical and the normative, but the technical argument is alluring because it offers the promise of objective correctness to trump messy compromises.<sup>21</sup> Instead, I suggest that these clever engineers are themselves underestimating policymakers and

corporate opponents: to stand on expertise instead of principle will work only if no other experts contradict you, and when those who wish to modify the network in undesirable ways don't like your engineering, they will simply buy new and better engineers. The use of technical arguments as a proxy for normative arguments produces a strange debate. If one argues that "the end-to-end principle renders the Internet an innovation commons" (Lessig 2001, 40), this is a debate in which control is ceded to whichever specialists are successful at defining the historical, or true Internet ("is this end-to-end, or is it not?"). In arguments from tradition, those that can define the past get to define the future.

The role of government. It is seductively easy to conceptualize technical, social, and legal mechanisms of control as different sorts of levers one can pull to steer the Internet (Clark et al. 2002). Writing in the end-to-end debate has pointed out the prevalence of technical control, and it has also called for more social (and even legal) control as a better "balance." In fact, what seems like "technical" control is nothing new. The technical, social, and legal always interpenetrate, and there is no way to guarantee technical control in the present Internet without some assertion of control, by an owner or by a government.

The process of network design should continue to include considerations of transparency, participation, and flexibility, but these should be explicit goals, and not pursued under the rubric of technical correctness or the end-to-end argument. Furthermore, the legitimate public policy role for governments lies not in protecting the Internet against those who would "break" it. This is merely a grant of authority to whomever is designated to interpret the Internet's fundamental nature and to write its history. Reflecting on the Internet's boon to innovation provides a logical

rationale for regulating transparency and participation. This is not a new role for government, even with respect to the Internet.

---

<sup>1</sup> First elaborated by Saltzer, Reed and Clark (1984).

<sup>2</sup> For an excellent review, see David (2001).

<sup>3</sup> This intriguing suggestion was made by David (2001) and also deemed impractical by him.

<sup>4</sup> The most visible popularizer of this notion may be Lessig (2001), the most precise Blumenthal and Clark (2001).

<sup>5</sup> After Internet “packet loss.”

<sup>6</sup> The historical details in this introduction are taken from Holtzmann & Pehrson (1995).

<sup>7</sup> This example is from Dvornik (1974).

<sup>8</sup> The simple beacon of light (fire) or smoke may predate the organized courier network, but the point here is the integration of the two.

<sup>9</sup> The details in this paragraph are taken from Lay (1992) and Forbes (1954).

<sup>10</sup> In telecommunications engineering, the notion of a single carrier owning all segments of a connection between two parties is also called “end-to-end,” but this use of the term is not relevant to this paper. This paper’s use of “end-to-end” refers to the definition used in computer networking from the 1970s forward.

<sup>11</sup> See, e.g., the section titled “Why the Computer Culture Matters” in Streeter (1999).

<sup>12</sup> The seven-layer model was developed as part of the Open Systems Interconnect (OSI) initiative of the International Organization for Standardization (ISO).

---

<sup>13</sup> Though the end-to-end concept appeared much earlier, along with the development of packet switching.

<sup>14</sup> The original example given in the article was the problem of “careful file transfer” (p. 510).

<sup>15</sup> Meaning that as an approach to designing the system, the design of the end is oriented to communication with the system (in this case, the switch) and not another endpoint. To simplify the term it might be easier to conceptualize it as “from end to system.”

<sup>16</sup> For instance, Integrated Services Digital Network (ISDN), a way for computers to communicate via telephone deployed in the 1980s, was an attempt to move some intelligence from the center of the (end-system) telephone network to the edge, and to implement some features in smarter terminals (called ISDN terminal adapters) than the traditionally dumb telephone.

<sup>17</sup> Stations in military courier networks like the post-house system were usually also distribution points for news and gossip and accepted non-military traffic when extra capacity was available. We can imagine, if we adopt today’s terms, a richer topography with non-uniform nodes, some broadcasting, some caching, and some filtering. All of this was, if not planned, at least known.

<sup>18</sup> “Users” may not necessarily mean novice users, however; some have suggested a base level of technical expertise is a prerequisite for a user to create a network innovation.

<sup>19</sup> See, for instance, the intriguing story of the Hush-a-Phone and the Carterfone (see Neuman, McKnight, & Solomon 1998, 176-178).

<sup>20</sup> e.g., Akamai.

---

<sup>21</sup> In this section I take issue with those who cover for normative goals with technical conclusions, but it is doubtful that the distinction between “technical” and “normative” is ever very useful. Every technological decision includes normative assumptions, even if these are usually so generally accepted or unexamined that they can be ignored.

## Sources Cited

Bar, François, Stephen Cohen, Peter Cowhey, J. Bradford DeLong, Michael Kleeman, and John Zysman. 2000. Access and innovation policy for the third-generation Internet.

*Telecommunications Policy*, 24(6/7): 489-518.

Bar, François and Annemarie Munk Riis. 2000. Tapping user-driven innovation: A new rationale for universal service. *The Information Society*, 16(2): 99-108.

Pinch, Trevor J. and Wiebe E. Bijker. 1987. The social construction of facts and artifacts: Or, how the sociology of science and the sociology of technology might benefit each other. In *The social construction of technological systems: New directions in the sociology and history of technology*, edited by Wiebe E. Bijker, Thomas P. Hughes, and Trevor Pinch. Cambridge: MIT Press: 17-50.

Blumenthal, Marjory S., and David D. Clark. 2001. Rethinking the design of the Internet: The end-to-end arguments vs. the brave new world. In *Communications policy in transition: The Internet and beyond*, edited by Benjamin M. Compaine and Shane Greenstein. Cambridge: MIT Press: 91-140.

Clark, David D. 1988. The design philosophy of the DARPA Internet protocols. *Computer Communication Review*, 18(4): 106-114.

Clark, David D., John Wroclawski, Karen R. Sollins, and Robert Braden. 2002, August 19. *Tussle in cyberspace: Defining tomorrow's Internet*. Pittsburgh: Paper presented at the

Annual Meeting of the ACM Special Internet Group on Data Communications (SIGCOMM).

David, Paul A. 2001. The evolving accidental information super-highway. *Oxford Review of Economic Policy*, 17(2): 159-187.

Dvornik, Francis. 1974. *Origins of intelligence services: The ancient Near East, Persia, Greece, Rome, Byzantium, the Arab Muslim Empires, the Mongol Empire, China, Muscovy*. New Brunswick: Rutgers University Press.

Forbes, Robert J. 1954. Roads to c. 1900. In *A history of technology*, vol. 4, edited by Charles Singer. Oxford: Clarendon Press.

von Hippel, Eric. (1988). *The sources of innovation*. Oxford: Oxford University Press.

Holzmann, Gerard J. and Björn Pehrson. 1995. *The early history of data networks*. Los Alamitos: IEEE Computer Society Press.

Isenberg, David S. 1997, August. The rise of the stupid network. *Computer Telephony*: 16-26.

Kruse, Hans, William Yurcik, and Lawrence Lessig. 2001. The InterNAT: Policy implications of the Internet architecture debate. In *Communications policy in transition: The Internet and beyond*, edited by Benjamin M. Compaine and Shane Greenstein. Cambridge: MIT Press: 141-158.

Lay, Maxwell G. 1992. *Ways of the world: A history of the world's roads and the vehicles that used them*. New Brunswick: Rutgers University Press.

- Lemley, Mark. A., and Lawrence Lessig. 2001. The end of end-to-end: Preserving the architecture of the Internet in the broadband era. *UCLA Law Review*, 48: 925.
- Lessig, Lawrence. 2000. Innovation, regulation, and the Internet. *The American Prospect*, 11(10).
- Lessig, Lawrence. 2001. *The future of ideas: The fate of the commons in a connected world*. New York: Random House.
- Mansell, Robin. 1994. *The new telecommunications: A political economy of network evolution*. Thousand Oaks: Sage.
- Moors, T. 2002. A critical review of “end-to-end arguments in system design,” In *Proceedings of the IEEE International Conference on Communications (ICC)*, vol. 5. New York: IEEE: 1214-9.
- Neuman, Russell W., Lee McKnight, and Richard Jay Solomon. 1998. *The Gordian knot: Political gridlock on the information highway*. Cambridge: MIT Press.
- Reed, David P. 2000. “The end of the end-to-end argument.”  
<http://www.reed.com/dprframeweb/dprframe.asp?section=paper&fn=endofendtoend.html>
- Saltzer, Jerome H., David P. Reed, and David D. Clark. (1984). End-to-end arguments in system design. *ACM Transactions on Computer Systems*, 2(4): 277-288.

Streeter, Thomas. 1999. "That deep romantic chasm": Libertarianism, neoliberalism, and the computer culture. In *Communication, citizenship, and social policy: Re-thinking the limits of the welfare state*, edited by A. Calabrese and J. C. Burgelman. New York: Rowman & Littlefield: 49-64.

## **Biographical Note**

Christian Sandvig is an Assistant Professor of Speech Communication at the University of Illinois at Urbana-Champaign and a Research Associate of the Centre for Socio-Legal Studies, Oxford University where he served as a Markle Foundation Information Policy Fellow from 2001-2002.

## **Acknowledgement**

The author would like to thank Helen Nissenbaum, Stephen Barley, Paul David, Ian “Gus” Hosein, William Drake, and Dieter Zinnbauer for their helpful suggestions. This research was kindly supported by the Markle Foundation Information Policy Fellowship at the Programme in Comparative Media Law and Policy at Oxford University, and by a Visiting Research Fellowship at the Oxford Internet Institute. An earlier version of this chapter that did not contain the arguments related to innovation was presented at “Computer Ethics: Philosophical Enquiries,” December 15, 2001, Lancaster, UK.

**TABLE 1.** *Networks Characterized by Design Strategy*

Ideal types:

*End-to-End Networks*

The earliest ad-hoc courier networks

Early non-switched (“party-line”) telephone networks

The Internet before widespread content caching

*End-System Networks*

The Chou Dynasty “post-house” courier network

Early switched telephone networks