Rethinking the design of the Internet: The end to end arguments vs. the brave new world

David D. Clark, M.I.T. Lab for Computer Science, ddc@lcs.mit.edu [1]

Marjory S. Blumenthal, Computer Science & Telecommunications Bd., mblument@nas.edu

Version for TPRC submission, August 10, 2000

Abstract

This paper looks at the Internet and the changing set of requirements for the Internet that are emerging as it becomes more commercial, more oriented towards the consumer, and used for a wider set of purposes. We discuss a set of principles that have guided the design of the Internet, called the *end to end arguments,* and we conclude that there is a risk that the range of new requirements now emerging could have the consequence of compromising the Internet's original design principles. Were this to happen, the Internet might lose some of its key features, in particular its ability to support new and unanticipated applications. We link this possible outcome to a number of trends: the rise of new stakeholders in the Internet, in particular Internet Service Providers; new government interests; the changing motivations of the growing user base; and the tension between the demand for trustworthy overall operation and the inability to trust the behavior of individual users.

Introduction

The end to end arguments are a set of design principles that characterize (among other things) how the Internet has been designed. These principles were first articulated in the early 1980s,[2] and they have served as an architectural model in countless design debates for almost 20 years. The end to end arguments concern how application requirements should be met in a system. When a general purpose system (for example, a network or an operating system) is built, and specific applications are then built using this system (for example, e-mail or the World Wide Web over the Internet), there is a question of how these specific applications and their required supporting services should be designed. The end to end arguments suggest that specific application-level functions usually cannot, and preferably should not, be built into the lower levels of the system—the core of the network. The reason why was stated as follows in the original paper:

The function in question can completely and correctly be implemented only with the knowledge and help of the application standing at the endpoints of the communications system. Therefore, providing that questioned function as a feature of the communications systems itself is not possible.

In the original paper, the primary example of this end to end reasoning about application functions is the assurance of accurate and reliable transfer of information across the network. Even if any one lower level subsystem, such as a network, tries hard to ensure reliability, data can be lost or corrupted after it leaves that subsystem. The ultimate check of correct execution has to be at the application level, at the endpoints of the transfer. There are many examples of this observation in practice.

Even if parts of an application-level function can potentially be implemented in the core of the network, the end to end arguments state that one should resist this approach if possible. There are a number of advantages of moving application-specific functions up out of the core of the network and providing only general-purpose system services there.

- The complexity of the core network is reduced, which reduces costs and facilitates future upgrades to the network.
- Generality in the network increases the chances that a new application can be added without having to change the core of the network.
- Applications do not have to depend on the successful implementation and operation of application-specific services in the network, which may increase their reliability.

Of course, the end to end arguments are not offered as an absolute. There are functions that can only be implemented in the core of the network, and issues of efficiency and performance may motivate core-located features. But the bias toward movement of function "up" from the core and "out" to the edge node has served very well as a central Internet design principle.

As a consequence of the end to end arguments, the Internet has evolved to have certain characteristics. The functions implemented "in" the Internet—by the routers that forward packets—have remained rather simple and general. The bulk of the functions that implement specific applications, such as e-mail, the World Wide Web, multi-player games, and so on, have been implemented in software on the computers attached to the "edge" of the Net. The edge-orientation for applications and comparative simplicity within the Internet together have facilitated the creation of new applications, and they are part of the context for innovation on the Internet.

Moving away from end to end

For its first 20 years, much of the Internet's design has been guided by the end to end arguments. To a large extent, the core of the network provides a very general data transfer service, which is used by all the different applications running over it. The individual applications have been designed in different ways, but mostly in ways that are sensitive to the advantages of the end to end design approach. However, over the last few years, a number of new requirements have emerged for the Internet and its applications. To certain stakeholders, these various new requirements might best be met through the addition of new mechanism in the core of the network. This perspective has, in turn, raised concerns among those who wish to preserve the benefits of the original Internet design.

Here are some (interrelated) examples of emerging requirements for the Internet of today:

Operation in an untrustworthy world: The examples in the original end to end paper assume that the end-points are in willing cooperation to achieve their goals. Today, there is less and less reason to believe that we can trust other end-points to behave as desired. The consequences of untrustworthy end-points on the Net include attacks on the network as a whole, attacks on individual end-points, undesired forms of interactions such as spam e-mail, and annoyances such as Web pages that vanish due to end-node aberrations.[3] The situation is a predictable consequence of dramatic growth in the population of connected people and its diversification to include people with a wider range of motivations for using the Internet, leading to uses that some have deemed misuses or abuses. Making the network more trustworthy, while the end-points cannot be trusted, seems to imply more mechanism in the center of the network to enforce "good" behavior.

Consider spam—unwanted bulk mail sent out for advertising or other purposes. Spam is not the most pernicious example of unwelcome end-node behavior—it usually annoys rather than disrupts. However, it provides a good example of how different approaches to control conform in different ways to the tenets of the end to end arguments. It is the person receiving spam, not the e-mail software, that desires to avoid receiving it. Staying within the end to end framework but applying the arguments at the ultimate end-point (the human using the system) implies that the sender sends the spam, the software at the receiver receives it, and then the human receiver deletes it. The underlying protocols, including both the TCP layer and the higher SMTP mail transfer layer, are just supporting mechanisms. However, because users resent the time (both personal and Internet-connection time) and sometimes money spent collecting and deleting the unwanted mail, some have proposed application-level functions elsewhere in the network, not just at the recipient's computer, to prevent spam from arriving at the edges.[4]

More demanding applications: The simple service model of the Internet (called "best effort delivery") makes no guarantee about the throughput that any particular application will achieve at any moment. Applications such as file transfer, Web access, or e-mail are tolerant of fluctuations in rate—while a user may be frustrated by a slow delivery, the application still "works." Today, a new set of applications is emerging, typified by streaming audio and video, that appear to demand a more sophisticated Internet service that can assure each data stream a specified throughput, an assurance that the best effort service cannot provide. Different approaches are possible, beginning with (re)design of applications to operate using only the current best effort service, perhaps by dynamically adjusting the fidelity of the transmitted information as the network throughput varies. At least some application designers reject this limitation on what they could design. Another approach would be to add new data transport services in the core of the network that provide predictable throughput and bounded delays, and there have been proposals along these lines.[5] However, the Internet Service Providers (see below) have not so far been willing to provide these new services. As a result, application builders have adopted the strategy of installing intermediate storage sites that position the streaming content close to the recipient, to increase the chance of successful delivery. Thus, unlike a simple end to end structure, the design of these new applications depends on a two-stage delivery via these intermediate servers.

ISP service differentiation: The deployment of enhanced delivery services for streaming media and other sorts of advanced Internet applications is shaped by the current business models of the larger Internet Service Providers. They (at least at present) seem to view enhanced data transport service as something to be provided within the bounds of the ISP as a competitive differentiator, sometimes tied to specific applications such as telephone service over the Internet, rather than a capability to be supported, end to end, across multiple providers' networks. If enhanced services are not provided end to end, then it is not possible to design applications needing these services using an end-point implementation. Thus, as discussed above, there is an acceleration in the deployment of applications based on intermediate servers that can be positioned within each ISP; content is delivered to ISP customers within the island of enhanced service. This approach has an additional effect that has aroused concern among consumer activists: the differentiation of applications generated by parties that can afford to promote and utilize ISP-specific intermediate servers from those that depend on potentially lower-performance, end to end transport.[6] The concern here, however, is that investment in closed islands of enhanced service, combined with investment in content servers within each island, decreases the motivation for investment in the alternative of open end to end services. Once started down one path of investment, the alternative may be harder to achieve.

The rise of third-party involvement: An increasingly visible issue is the demand by third parties to interpose themselves between communicating end-points, irrespective of the desires of the ends.[7] Third parties may include officials of organizations (e.g., corporate network or ISP administrators implementing organizational policies or other oversight) or officials of governments, whose interests may range from taxation to law enforcement and public safety. Court-ordered wiretaps illustrate government interposition as a third party, whereas mandatory blocking of certain content may involve either government or organizational interposition.

Less sophisticated users: The Internet was designed, and used initially, by technologists. As the base of users broadens, the motivation grows to make the network easier to use. By implying that substantial software is present at the end-node, the end to end arguments are a source of complexity to the user: that software must be installed, configured, upgraded, and maintained. It is much more appealing to some to take advantage of software that is installed on a server somewhere else on the network.[8] The importance of ease of use will only grow with the changing nature of consumer computing. The computing world today includes more than PCs. It has embedded processors, portable user-interface devices such as computing appliances or personal digital assistants (PDAs, such as Palm devices), Web-enabled televisions and advanced set-top boxes, new kinds of cell-phones, and so on. If the consumer is required to set up and configure separately each networked device he owns, what is the chance that at least one of them will be configured incorrectly? That risk would be lower with delegation of configuration, protection, and control to a common point, which can act as an agent for a pool of devices.[9] This common point would become a part of the application execution context. With this approach, there would no longer be a single indivisible end-point where the application runs.

While no one of these trends is by itself powerful enough to transform the Internet from an end to end network to a network with centralized function, the fact that they all might motivate a shift in the same direction could herald a significant overall change in the shape of the Net. Such change would alter the Internet's economic and social impacts. That recognition lies behind the politics of those changes and the rhetoric of parties for and against various directions that might be taken in developing and deploying mechanisms. That the end to end arguments have recently been invoked explicitly in political debates reflects the growth in the stakes and the intensification of the debates.[10] At issue is the conventional understanding of the "Internet philosophy": freedom of action, user empowerment, end-user responsibility for actions undertaken, and lack of controls "in" the Net that limit or regulate what users can do. The end to end arguments fostered that philosophy because they enabled the freedom to innovate, install new software at will, and run applications of the user's choice.

The end to end arguments presuppose to some extent certain kinds of relationships: between communicating parties at the ends, between parties at the ends and the providers of their network/Internet service, and of either end users or ISPs with a range of third parties that might take an interest in either of the first two types of relationship (and therefore the fact or content of communications). In cases where there is a tension among the interests of the parties, our thinking about the objectives (and about the merit of technical mechanisms we might or might not add to the network) is very much shaped by our values concerning the specifics of the case. If the communicating parties are described as "dissidents," and the third party trying to wiretap or block the conversation is a "repressive" government, most people raised in the context of free speech will align their interests with the end parties. Replace the word "dissident" with "terrorist," and the situation becomes less clear to many. Similarly, when are actions of an ISP responsible management of its facilities and service offerings, and when are they manipulative control of the nature and effective pricing of content and applications accessed through its facilities and services?

Perhaps the most contentious set of issues surrounds the increasing third-party involvement in communication between cooperating users. When communicating end-points want to communicate, but some third party demands to interpose itself into the path without their agreement, the end to end arguments do not provide an obvious framework to reason about this situation. We must abandon the end to end arguments, reject the demand of a third party because it does not "fit" our technical design principles, or find another design approach that preserves the power of the end to end arguments as much as possible.

Preservation of the end to end arguments would imply that if, in a given jurisdiction, there are political or managerial goals to be met, meeting them should be supported by technology and policies at higher levels of the system of network-based technology, not by mechanism "in" the network. The new context of the Internet implies that decisions about where to place mechanisms will be more politicized and that more people may need more convincing about the merits of a pro-end to end decision than in the Internet's early days. It is time for a systematic examination of what it means to uphold or deviate from the end to end arguments as the Internet evolves.

The rest of this paper is organized as follows. We first catalog a number of new requirements for controls and protections in today's communication. We document the emerging calls for the Internet to address these new requirements. We then identify a range of possible solutions that might be used to meet these requirements. We look at technical options, but we emphasize that non-technical approaches (legal, social, economic) are important, valid, and often preferable. We then look at the implications for the rights and responsibilities of the various parties that comprise the Internet—the consumer as user, the commercial ISPs, the institutional network providers, governments, and so on. We describe the range of emerging players, to emphasize the complexity of the space of stakeholders in this new world. We conclude by offering some observations and speculations on what the most fundamental changes are and what is most important to preserve from the past.

Examples of requirements in today's communication

As the previous section suggested, many of the complexities in communication today reflect more diverse patterns of interaction among the different players. This section catalogs a number of requirements, to illustrate the breadth of the issues and to suggest the range of solutions that will be required.

Users communicate but don't totally trust each other

One important category of interaction occurs when two (or more) end-nodes want to communicate with each other but do not totally trust each other. There are many examples of this situation:

- Two parties want to negotiate a binding contract: they may need symmetric proof of signing, protection from repudiation of the contract, and so on.[11]
- One party needs external confirmation of who the other party in the communication is.
- At the other extreme, two parties want to communicate with each other but at least one of the parties wants to preserve its anonymity. This topic is of sufficient importance that we consider it in detail below.

Users communicate but desire anonymity

There are a number of circumstances in which a desire for anonymity might arise, from anonymous political speech and whistle blowers to reserving one's privacy while looking at a Web site. At least in the United States, the privilege of anonymous public political speech is seen as a protected right. In this context, the speakers will seek assurance that their anonymity cannot be penetrated, either at the time or afterwards. This concern is directed at third parties—not only individuals who might seek to uncover the speaker, but the government itself, which might want to repress certain expressions. Another example is on-line voting. Individual voters need some external assurance that their votes are anonymous. The voting system needs to ensure that only registered voters can vote and each votes at most once. The citizens, collectively, seek assurance that voting is not disrupted by some denial of service attack, the vote tally is accurate, and that there is no opportunity for voting fraud. A third example is the call for anonymous electronic cash on the Internet so that one could complete an online purchase anonymously.[12]

The desire for anonymity is an example of a situation where the interests of the different end-parties may not align. One end may wish to hide its identity, while the other end may need that identity or at least to confirm some attributes (e.g., status as an adult, or citizenship) in order to authorize some action.

One's identity can be tracked on the network in a number of ways. For example, low level identification such as e-mail addresses or the IP address of the user's computer can be used to correlate successive actions and build a user profile that can, in turn, be linked to higher-level identification that the user provides in specific circumstances.[13] The dynamic interplay of controls (e.g., attempts to identify) and their avoidance is an indication that the Internet is still flexible, the rules are still evolving, and the final form is not at all clear.[14]

End parties do not trust their own software and hardware

There is a growing perception that the hardware and software that are available to consumers today behave as a sort of double agent, releasing information about the consumer to other parties in support of marketing goals such as building profiles of individual consumers. For example, Web browsers today store "cookies" (small fragments of information sent over the network from a Web server) and send that data back to the same or different servers to provide a trail that links successive transactions, thereby providing a history of the user's behavior.[15] Processors may contain unique identifiers that can distinguish one computer from another, and various programs such as browsers could be modified to include that identifier in messages going out over the Internet, allowing those messages to be correlated.[16] Local network interfaces (e.g., Ethernet) contain unique identifiers, and there is fear that those identifiers might be used as a way to keep track of the behavior of individual people.[17] These various actions are being carried out by software (on the user's computer) that the user is more or less required to use (one of a small number of popular operating systems, Web browsers, and so on) as well as elective applications.[18]

The ends vs. the middle: third parties assert their right to be included in certain sorts of transactions

Another broad class of problem can be characterized as a third party asserting its right to interpose itself into a communication between end-nodes that fully trust each other and consider themselves fully equipped to accomplish their communication on their own. There are many examples of this situation.

- Governments assert their right to wiretap (under circumstances they specify) to eavesdrop on certain communications within their jurisdiction.

- Governments, by tradition if not by explicit declaration of privilege, spy on the communications of parties outside their jurisdiction.

- Governments take on themselves the right to control the access of certain parties to certain material. This can range from preventing minors from obtaining pornographic material to preventing citizens from circulating material considered seditious or unwelcome by that government.

- Governments assert their right to participate in specific actions undertaken by their citizens for public policy reasons, such as enforcement of taxation of commercial transactions.

- Private ISPs assert their right to regulate traffic on their networks in the interests of managing load, and in order to segregate users with different intentions (e.g., those who provide or only use certain application services), in order to charge them different amounts.

- Private organizations assert their right to control who gets access to their intranets and to their gateways to the Internet, and for what purposes.

- Private parties assert their right to intervene in certain actions across the network to protect their rights (e.g., copyright) in the material being transferred.

The requirements of private parties such as rights holders may be as complex as those of governments. The end to end arguments, applied in a simple way, would suggest that a willing sender can use any software he chooses to transfer material to willing receivers. The holders of intellectual property rights may assert that, somewhat like a tax collector but in the private domain, they have the right to interpose themselves into that transfer to protect their rights in the material (and ability to collect fees), which thus potentially becomes a network issue.[19]

For each of these objectives, there are two perspectives: There are mechanisms that the third parties use to inject themselves into the communication, and there are actions that the end-parties use to try to avoid this intervention. In general, mechanisms with both goals can be found inside networks, representing a dynamic, evolving balance of power between the parties in question.

Different third-party objectives trigger a range of requirements to observe and process the traffic passing through the network. Some objectives, such as certain forms of wiretapping, call for access to the complete contents of the communication. On the other hand, some objectives can be met by looking only at the IP addresses and other high-level identifying information describing the communication. These latter activities, referred to as *traffic analysis*, are common in the communications security and law enforcement communities, where they may be regarded as second-best compared to full-content access.

In the contemporary environment, attention to communications patterns extends beyond the government to various private parties, in part because technology makes it possible. A kind of traffic analysis is appearing in the context of large, organizational users of the Internet, where management is policing how organizational resources are used (e.g., by monitoring e-mail patterns or access to pornographic Web sites[20]). Finally, ISPs may use traffic analysis in support of their traffic engineering. ISPs have asserted that it is important for them to examine the traffic they are carrying in order to understand changing patterns of user behavior; with that information they can predict rates of growth in different applications and thus the need for new servers, more network capacity, and so on. The rise of high-volume MP3 file exchanges, boosted by Napster (a directory of individual collections) and Gnutella for peer-to-peer sharing, illustrates the sort of phenomenon that ISPs need to track. Normally, they do not need to look at the actual data in messages, but only at the identifiers that indicate which application is being used (e.g., whether a message is e-mail or a Web access).

The desire by some third party to observe the content of messages raises questions about the balance of power between the end-points and the third party. As we detail below, an end-point may try to prevent any observation of its data, in response to which the third party may try to regulate the degree to which the end-points can use such approaches. There may be other points on the spectrum between total privacy and total accessibility of information, for example *labels* on information that interpret it or reveal specific facts about it. Labeling of information is discussed below.

One party tries to force interaction on another

The example of asymmetric expectations among the end-nodes reaches its extreme when one party does not want to interact at all, and the other party wishes to force some involvement on it. This network equivalent of screaming at someone takes many forms, ranging from application-level flooding with unwanted material (e.g., e-mail spam) to what are seen as security attacks: penetration of computers with malicious intent (secretly, as with Trojan horses, discussed below, or overtly), or the anti-interaction problem of denial of service attacks, which can serve to prevent any interactions or target certain kinds.[21]

Even when a user is communicating with a site that is presumed harmless, there are always risks of malicious behavior—classic security breaches and attacks, deception and misdirection of the user, transmittal of viruses and other malicious code, and other snares.[22] The classic end to end arguments would say that each end-node is responsible for protecting itself from attacks by others (hence the popularity of anti-virus software), but this may not be viewed as sufficient control in today's complex network.

One classic computer security attack is the so-called Trojan horse, in which a user is persuaded to install and use some piece of software that, while superficially performing a useful task, is in fact a hostile agent that secretly exports private information or performs some other sort of clandestine and undesirable task affecting the recipient's system and/or data. It is not clear how often Trojan horse programs actually succeed in achieving serious security breaches, but there is growing concern that "trusting" browsers may be blind to Trojan horses that can be deposited on end-systems through interactions with server software designed with malicious intent.[23]

Multiway communication

The examples above are all cast in the framework of two-party communication. But much of what happens on the Internet, as in the real world, is multi-party. Any public or semi-public network offering has a multiway character. Some interactions, like the current Web, use a number of separate two-party communications as a low-level technical means to implement the interaction from a server to multiple users. Others, like teleconferencing or receiving Internet-based broadcast material (audio or video), may also involve multiway communication at the network level, traditionally called multicast.

Part of what makes multiway applications more complex to design is that the multiple end-points may not function equally. Different participants may choose to play different roles in the multiway interaction, with different degrees of trust, competence, and reliability. Some will want to participate correctly, but others may attempt to disrupt the communication. Some may implement the protocols correctly, while others may crash or malfunction. These realities must be taken into account in deciding how to design the application and where functions should be located.

In general, in a two-party interaction, if one end seems to be failing or malicious, the first line of defense is to terminate the interaction and cease to communicate with that party. However, in a multiway communication, it is not acceptable for one broken end-point to halt the whole interaction. The application must be designed so that it can distinguish between acceptable and malicious traffic and selectively ignore the latter. It may be possible to do this within the end-node, but in other cases (e.g., where the network is being clogged by unwanted traffic) it may be necessary to block some traffic inside the network. This will require the ability to install traffic filters inside the network that are specific as to source address and application type as well as multicast destination address.

Summary—what do these examples really imply?

This set of examples is intended to illustrate the richness of the objectives that elements of society may desire to impose on its network-based communication. The existence or identification of such examples does not imply that all of these goals will be accepted and reflected in new technical mechanisms (let alone judgment of their merits). Rather, it shows that the world is becoming more complex than it was when the simple examples used to illustrate the end to end arguments were articulated.

Does this mean that we have to abandon the end to end arguments? No, it does not. What is needed is a set of principles that interoperate with each other—some build on the end to end model, and some on a new model of network-centered function. In evolving that set of principles, it is important to remember that, from the beginning, the end to end arguments revolved around requirements that could be implemented correctly at the end-points; if implementation inside the network is the only way to accomplish the requirement, then an end to end argument isn't appropriate in the first place.[24] The end to end arguments are no more "validated" by the belief in end-user empowerment than they are "invalidated" by a call for a more complex mix of high-level functional objectives.

Technical responses

The preceding section catalogued objectives that have been called for (in at least some quarters) in the global Internet of tomorrow. There are a number of ways that these objectives might be met. In this section, we examine technical responses that have been put forward and organize them into broad categories.

The different forms of the end to end arguments

The end to end arguments apply at (at least) two levels within the network. One version applies to the core of the network—that part of the Internet implemented in the routers themselves, which provide the basic data forwarding service. Another version applies to the design of applications.

The end to end argument relating to the core of the network claims that one should avoid putting application-specific functions "in" the network, but should push them "up and out" to devices that are attached "on" the network. Network designers make a strong distinction between two sorts of elements—those that are "in" the network and those that are "attached to," or "on," the network. A failure of a device that is "in" the network can crash the network, not just certain applications; its impact is more universal. The end to end argument at this level thus states that services that are “in” the network are undesirable because they constrain application behavior and add complexity and risk to the core. Services that are “on” the network, and which are put in place to serve the needs of an application, are not as much of an issue because their impact is narrower.

From the perspective of the core network, all devices and services that are attached to the network represent end-points. It does not matter where they are—at the site of the end user, at the facilities of an Internet Service Provider, and so on. But when each application is designed, an end to end argument can be employed to decide where application-level services themselves should be attached. Some applications have a very simple end to end structure, in which computers at each end send data directly to each other. Other applications may emerge with a more complex structure, with servers that intermediate the flow of data between the end-users. For example, e-mail in the Internet does not normally flow in one step from sender to receiver. Instead, the sender deposits the mail in a mail server, and the recipient picks it up later.

Modify the end-node

The approach that represents the most direct lineage from the Internet roots is to try to meet new objectives by modification of the end-node. In some cases, placement of function at the edge of the network may compromise performance, but the functional objective can be met. If spam is deleted before reaching the recipient or afterwards, it is equally deleted. The major different is the use of resources—network capacity and user time—and therefore the distribution of costs—with deletion before or after delivery. The difference, in other words, is performance and not “correctness” of the action.

In other cases, implementation in the end-node may represent an imperfect but acceptable solution. Taxation of transactions made using the Internet[25] is a possible example. Consider an approach that requires browser manufacturers to modify their products so that they recognize and track taxable transactions. While some people might obtain and use modified browsers that would omit that step, there would be difficulties in obtaining (or using) such a program, especially if distributing (or using) it were illegal. One approach would be to assess the actual level of non-compliance with the taxation requirement, make a judgment as to whether the level of loss is acceptable, and develop complementary mechanisms (e.g., laws) to maximize compliance and contain the loss.[26] As we discuss below, a recognition that different end-points play different roles in society (e.g., a corporation vs. a private citizen) may make end-located solutions more robust and practical.

Control of access to pornography by minors is another example of a problem that might be solved at an end-point, depending on whether the result is considered robust enough. One could imagine that objectionable material is somehow labeled in a reliable manner, and browsers are enhanced to check these labels and refuse to retrieve the material unless the person controlling the computer (presumably an adult) has authorized it. Alternatively, if the user does not have credentials that assert that he or she is an adult, the server at the other end of the connection can refuse to send the material.[27] Would this be adequate? Some minors might bypass the controls in the browser. Adventurous teenagers have been bypassing controls and using inaccurate (including forged or stolen) identification materials for a long time, and it is hard to guarantee that the person using a given end-system is who he or she claims to be. These outcomes represent leakage in the system, another case where compliance is less than one hundred percent. Is that outcome acceptable, or is a more robust system required?

In other circumstances, it would seem fruitless to depend on end-node modification. As the 1990s debates about government-accessible encryption keys illustrate, if the goal is to eavesdrop on suspected terrorists, there is no way to compel them to use only law-abiding software (a clear illustration of the end to end argument that the end-nodes may do as they please in carrying out a transaction). Even if some terrorists communicate "in the clear," it does not give much comfort to law enforcement if there is one encrypted conversation in particular that it wants to listen in on.

Adding functions to the core of the network

Examination of some emerging network requirements has led to a call for new mechanism "in" the network, at the level of the routers that forward packets across the Internet. This outcome is the most explicit challenge to the end to end arguments, because it puts function into the network that may prevent certain applications from being realized.

There is an important difference between the arguments being made today for function in the network and arguments from the past. In the past, the typical proposal for network-level function had the goal of trying to help with the implementation of an application. Now, the proposals are as likely to be hostile as helpful—addition of mechanism that keeps things from happening, blocks certain applications and so on.

Here are a number of examples where this approach is already being adapted today; others are contemplated.[28]

Firewalls: The most obvious example of a node inserted into the Internet today is a security firewall used to protect some part of the network (e.g., a corporate region) from the rest of the Internet. Firewalls inspect passing network traffic and reject communications that are suspected of being a security threat.

Traffic filters: Elements such as firewalls can perform tasks beyond providing protection from outside security attacks. They can affect traffic in both directions, so they can be programmed to prevent use of some applications (e.g., game playing) or access to inappropriate material (e.g., known pornography sites), as well as a number of other functions. Traffic filters can thus become a more general tool for control of network use.

Network address translation elements: Today, devices called Network Address Translation (NAT) boxes are being used in the Internet to deal with the shortage of Internet addresses and to simplify address space management.[29] By modifying the IP addresses in the packets, they may contribute to protecting user identity from other end-points. These are sometimes integrated in with firewall functions—e.g., as a part of their operation they can limit the sorts of applications that are permitted to operate. NAT boxes are usually installed by managers of organizational networks and some ISPs. There have also been proposals to use address translation on a larger scale, perhaps for an entire country, as a way to control access into and out of that country.

However, the deployment of NAT requires many adjustments elsewhere. An original design principle of the Internet is that IP addresses are carried unchanged end to end, from source to destination across the network. The next level protocol normally used above IP, TCP, verifies this fact. With the introduction of NAT boxes, which rewrite the IP addresses in packets entering or leaving a region of the network, these boxes also had to modify the information sent at the TCP level; otherwise, the TCP error checking would have reported an addressing error. The more difficult problem is that some higher level protocols (e.g., applications) also make use of the IP address; this implies that for the NAT box to preserve correct operation, it must understand the design of specific applications, a clear violation of the end to end arguments.

Finally, IP addresses are used in additional ways in practice. For example, some site licenses for software use the IP address of the client to control whether to give the client access to the server. Changing the apparent address of the client can cause this sort of scheme to malfunction.

Design issues in adding mechanism to the core of the network

There are two issues with any control point imposed "in" the network. First, the stream of data must be routed through the device, and second, the device must have some ability to see what sort of information is in the stream, so that it can make the proper processing decisions.

Imposing a control element into the path of communication

Packets flowing from a source to a destination can take a variety of paths across the Internet, since the best routing options are recomputed dynamically while the Internet is in operation. There is no single place in the Internet where a control point can be interposed in an unspecified flow. However, for a known flow, with a given source or destination, there is often an accessible location at which to insert a control point. For most users, access to the Internet is over a single connection, and a control point could be associated with that link. A corporation or other large user normally has only a small number of paths that connect it into the rest of the Internet, and these paths provide a means to get at the traffic from that organization. It is this topological feature that provides a place for an organization to install a firewall. The point where this path connects to an ISP similarly provides a means to monitor the traffic. Thus, the government could implement a wiretap order by instructing the ISP servicing the user to install a control point where the party in question attaches to it—a tack that has been attempted.[30]

Once the traffic has entered the interior of the public Internet, it becomes much more difficult to track and monitor. Thus, the ISP that provides initial access for a user to the Internet will, as a practical matter, play a special role in any mandated imposition of a monitoring device on a user.[31] As governments take increasing interest in what is being transmitted over the Internet, we can expect that the ISPs that provide the point of access for users to the Internet will be attractive to governments as vehicles for implementing certain kinds of controls associated with public policy objectives.[32]

Revealing or hiding the content of messages

Assuming that the network routing problem has been solved, and the traffic to be monitored is passing through the control point, the other issue is what aspects of the information are visible to the control device. There is a spectrum of options, from totally visible to totally masked. A simple application of the end to end arguments would state that the sender and receiver are free to pick whatever format for their communication best suits their needs. In particular, they should be free to use a private format, encrypt their communications, or use whatever means they choose to keep them private. Encryption can be the most robust tool for those who want to protect their messages from observation or modification. When strong encryption is properly implemented, the control device can only look at source and destination IP addresses, and perhaps other control fields in the packet header. As discussed above, traffic analysis is the only form of analysis possible in this case.

The goal of end to end privacy is in direct conflict with the goal of any third party that desires to take some action based on the content of the stream. Whether the goal is to tax an e-commerce transaction, collect a fee for performance of copyrighted music, or filter out objectionable material, if the nature of the contents is completely hidden, there is little the intermediate node can do, other than to block the communication all together. This situation could lead to a requirement that the device be able to see and recognize the complete information. Either the outcome of total privacy or total disclosure of content may be called for in specific cases, but it is valuable to identify possible compromises.

Labels on information

One way to reveal some information about the content of a message without revealing the content itself is to label the message. Labels, which would be visible in the network, represent one possible compromise between the rights of the end-node parties to transmit anything they want, perhaps encrypted for privacy, and the rights of some third party to observe or act on what is sent. Labels also represent a way to augment the actual information in the message, for example to impose a simple framework of content types on arbitrary application data. For example, a wide range of messages can be described with the simple label, “Advertising.” California law requires that all unsolicited advertising e-mail have “ADV:” at the beginning of the subject.[33] There is an important duality in the potential use of labels: they could be used to identify both content and users. For example, the transfer of pornographic material might be required to be labeled as “objectionable for a minor,” while the request for that material might carry the label of the class of person requesting it. Which scheme is used may depend on where the trust lies, and who can be held accountable.[34] Almost of necessity, such labeling schemes will be criticized as lacking generality and expressivity and as constraining all parties in some ways, especially for qualities that go beyond the factual. Labeling places a burden on the content producer or other party to attach accurate labels, and the question becomes whether this requirement is enforceable.[35]

As a practical matter, labels may become commonplace anyway in U.S. commercial communications, as the Federal Trade Commission moves to extend practices and policies associated with preventing deception in conventional media (which have led to the convention of labeling advertisement as such, for example) to the Internet.[36] Also, data labeling is a key building block of many filtering schemes, and it allows the filtering to be done both inside and at the edge of the network.

Labeling schemes side-step the practical problem of building an intermediate node that can analyze a message and figure out what it means. One could imagine writing a program that looks at the text of mail and concludes that it is bulk advertising, or looks at images and concludes that they are objectionable, or looks at a Web transfer and concludes that it is an online purchase. Although concepts for such programs are being pursued, they raise many troublesome issues, from the reliability of such controls to the acceptability of casting the decision-making in the form of a program in the first place.

There are several proposals for use of labels as a middle point on a spectrum of content visibility, although there are few used in practice today. One of the more visible label schemes in the Internet today is the Platform for Internet Content Selection (PICS) standard for content labeling,[37] which was developed by the World Wide Web Consortium as an approach to identification of potentially objectionable material. The PICS standard is a powerful approach to content labeling, since it permits content to be labeled by third parties as well as the content producers. This generality permits different users of content with different goals and values to subscribe to labeling services that match their needs. The label is not attached to the page as it is transferred across the network, but it is retrieved from the labeling service based on the page being fetched. The content can be blocked either in the end-node (an end to end solution) or in an application-level relay, specifically a Web proxy server (an in-the-net solution).[38] While PICS has many interesting and useful features, it has also attracted its share of criticism, most vocally the concern that the "voluntary" nature of the PICS labels could become mandatory in practice under government pressure. PICS might thus end up as a tool of government censorship.[39] This concern would seem to apply to any scheme for labels that can be observed in the network. Labeling schemes should not be seen as a panacea for all content issues, but they are a mid-point on a spectrum between lack of any visibility of what is being carried and explicit review and regulation of content.

Another example of content labels today are the metadata tags that are found on Web pages.[40] These are being used to help guide search engines in their cataloging of pages. Metadata tags can include keywords that do not actually appear in the visible part of the page; this feature can either be used to solve specific cataloging problems, or to promote a page to the top of a list of search results. As of today, these labels are not used for control inside the net but only for lookup, and they illustrate some of the problems with the use of labels.[41]

The Internet today provides a minimal label on most communications, the so-called "port number," which identifies which application at the end-point the message is intended for—Web, e-mail, file transfer, and so on. These numbers can be used to classify the packets crudely, and this ability is used today in a number of ways. ISPs and institutional network managers observe the port numbers to build models of user behavior to predict changes in demand. In some cases, they also refuse to forward traffic to and from certain port numbers, based on the service contract with the user. Some application developers have responded by moving away from predictable port numbers.

Design of applications—the end to end argument at a higher level

The previous discussion concerned augmentation of the core of the network with new sorts of functions, which in the current world are more concerned with control and filtering than with enhancing application. We now look at the design of the applications themselves. There are two trends that can be identified today. One is the desire on the part of different parties, either end-users or network operators, to insert some sort of server into the data path of an application that was not initially designed with this structure. This desire may derive from goals as diverse as privacy and performance enhancement. The other trend is that application requirements are becoming more complex, which sometimes leads away from a simple end to end design and toward the use of additional components as a part of the application.

Here are some examples of application-level services that are being employed today to augment or modify application behavior.

Anonymizing message forwarders: One strategy for users to achieve anonymity and to protect their communications from third party observation is to use a third-party service and route traffic through it so that possible identification in the messages can be removed. Services that make Web browsing anonymous are popular today,[42] and services with the specific goal of preventing traffic analysis are available.[43] Anonymous mail relays include simple remailers and more complex systems such as the nym server.[44] To use these devices, the end-node constructs the route through one (or usually more) of them to achieve the desired function. It is critical that the user construct the route, because preserving anonymity depends on the data following a path among the boxes that only the user knows; the ISP, for example, or any other third party should not be able to determine the path directly. Careful use of encryption is employed in these schemes to hide the route as well as the identity from unwanted observation.[45]

Helpful content filtering: The mail servers in use today can, in principle, be used to perform filtering and related processing on mail. Since the mail is routed through these devices anyway, server-filtering provides an option to remove spam or other objectionable material before it is even transferred to the receiving host.[46] Filtering can be done in a number of ways, consistent with the spectrum of access to content discussed above: looking at labels on the mail, matching of sender against a list of acceptable correspondents, or processing the content of the message (e.g., to detect viruses).

Content caches: The World Wide Web, perhaps the most visible of Internet applications today, was initially designed with a simple, two-party end to end structure. However, if a number of users fetch the same popular Web page, the original design implied that the page would be fetched from the server over and over again, and transferred multiple times across the network. This observation led to the suggestion that when a page was sent from a server to a user, a copy be made and "cached" at a point near the user, so that if a nearby user requested the page a second time, this subsequent request could be satisfied with the cached copy. Doing so may offer some significant performance advantages, but it does break the end to end nature of the Web; for example the server can no longer tell how many times its pages have been retrieved, nor can the server perform user-specific actions such as advertisment placement.[47]

More complex application design—using trusted third parties

Many issues in application design today derive in some way from a lack of trust between the users that are party to the application. A fundamental approach is to use a mutually trusted third party located somewhere on the network to create a context in which a two-party transaction can be successfully carried out.[48] In other words, what might have been a simple two-party transaction, conforming to the end to end arguments in a straightforward way, becomes a sequence of interactions among the three or more parties. Each interaction is nominally end to end (these third parties need not be "in" the network), but its robustness depends on the larger context composed of the whole sequence.

Some simple examples of what a trusted third party might do include signing and date-stamping of messages (even if a message is encrypted, an independent signature can provide protection from some forms of repudiation) or assuring simultaneous release of a message to multiple parties.[49] Another class of trusted third party will actually examine the content of messages and verify that the transaction is in proper form. This role is somewhat analogous to that of a notary public.[50]

Another role of a third party is to provide credentials that serve to give each party in a transaction more assurance as to the identity, role, or level of trustworthiness of the other party. Examples include voter registration, certification of majority (e.g., to permit access to material deemed harmful to minors) and so on. This role of the third party relates to the labeling both of content and users. It may be that a third party is the source of labels that are used to classify material, as discussed above in the context of PICS. There are other forms of tokens, beyond credentials that describe users and content, that can be obtained in advance. For example, anonymous electronic cash from a trusted third party (analogous to a bank) provides a context in which two-party anonymous purchase and sale can be carried out.

Public-key certificates

An important role for a third party occurs when public key cryptography is used for user authentication and protected communication. A user can create a public key and give it to others, to enable communication with that user in a protected manner. Transactions based on a well-known public key can be rather simple two-party interactions that fit well within the end to end paradigm. However, there is a key role for a third party, which is to issue a Public Key

Certificate and manage the stock of such certificates; such parties are called certificate authorities. The certificate is an assertion by that (presumably trustworthy) third party that the indicated public key actually goes with the particular user. These certificates are principal components of essentially all public key schemes, except those that are so small in scale that the users can communicate their public keys to each other one to one, in an ad hoc way that is mutually trustworthy.

The act of obtaining the certificate can be done in advance. In most schemes, there is also a step that has to be done after a transaction; this step is tricky in practice. It can happen that a user loses his private key (the value that goes with the given public key) through inadvertence or theft; alternatively, a user may become unworthy in some way relevant to the purpose for which the certificate has been issued. Under such circumstances, the certificate authority (third party) would want to revoke the certificate. How can this be known? The obvious (and costly) approach is for any party encountering a public key certificate to contact the third party that issued it to ask if it is still valid. Although that kind of interaction is seen commonly with electronic credit-card authorization, the potential for more uses of certificates and more users poses the risk of a substantial performance burden on the certifying authority, because it would end up receiving a query every time any of its certificates is used in a nominally two-party transaction and because there are inherent lags in the sequence of events leading to revocation. As a result, it is possible that the complexity may far exceed that associated with, say, invalid credit-card authorization today. There have been proposals to improve the performance implications of this revocation process, the details of which do not matter. But a general point emerges: Either the recipient of a public key certificate checks it in “real time,” during the process of a transaction with the party associated with that key, or it completes the transaction and then later verifies the status of the party in question, with the risk that the transaction already completed is not appropriate.[51]

In general, in a complex transaction involving multiple parties, there is an issue concerning the timing of the various actions by the parties. Voter registration does not happen at the time of voting, but in advance. However, unless there is periodic checking, one can discover that deceased voters are still voting, as well as voters that have just left town and registered elsewhere. A PICS rating of a page is necessarily done in advance. Even if the PICS rating is checked in real time as the page is retrieved, the rating itself may be out of date because the content of the page has changed. A generalization that often seems to apply is that the greater in time the difference between the preliminary or subsequent interaction with the third party and the transaction itself, the greater the risk that the role played by the third party is less reliable.

The larger context

It is important to consider the larger context in which these technical mechanisms exist. That context includes the legal and social structure of the economy, the growing motivations for trustworthiness, and the fact that technology, law, social norms, and markets combine to achieve a balance of power among parties.

Non technical solutions: the role of law in cyberspace

Just because a problem arises in the context of a technical system such as the Internet, it is not necessary that the solution be only technical.[52] In fact, the use of law and other non-technical mechanisms could be seen as consistent with the end to end arguments at the highest level—functions are moved “up and out,” not only from the core of the network but from the application layer as well, and positioned outside the network all together.

For example, to control the unwanted delivery of material to fax machines (spam in the fax world) there are laws that prohibit certain sorts of unsolicited fax transmissions and require that a sending fax machine attach its phone number so that the sender can be identified.[53] Similarly, the growth of computer-based crime has led to criminalization of certain behavior on the Internet: the 1987 Computer Security Act focused on “federal-interest” computers, and, thanks in large part to the proliferating use of the Internet and the associated tendency for computers to be networked, throughout the 1990s there was growing law enforcement attention, and legislation, relating to abuses of computers in both private and public sectors.[54]

The proliferation of labeling schemes points to the interplay of technical and legal approaches. The network can check the labels, but enforcement that the labels are accurate may fall to the legal domain.[55] This, of course, is the case in a variety of consumer protection and public safety situations; for example, the Federal Trade Commission regulates advertising—including claims and endorsement—in ways that affect content and format generally, and it has begun to examine the need for regulation relating to on-line privacy protection, while the Securities and Exchange Commission regulates financial claims, and the Food and Drug Administration regulates claims relating to food, pharmaceuticals, and medical devices. The FTC and others recognize that labels are an imperfect mechanism, in that people may ignore them, they may not apply to foreign sources, and they are subject to legal constraints in the United States as compelled speech, but labeling constitutes less interference with the market than, say, outright banning of products that raise policy concerns.

To date, on the Internet, enforcement has been less formal. The situation is similar to others, where voluntary action by industry may yield “self-regulation” of label content intended to avoid or forestall government regulation; content ratings for motion pictures, television shows (now associated with the V-chip[56]), and computer games provide examples that have attracted both public and governmental scrutiny; more entrepreneurial examples include the quality labeling emerging for Web sites from the Better Business Bureau and new entities that have arisen for this purpose. In other cases, a more popular vigilantism may be invoked: as the daily news have shown in reporting public outcry against companies misusing personal information (e.g., Amazon.com, RealNetworks, or DoubleClick),[57] public scrutiny and concern itself can have an impact.[58] Overall, mechanisms outside of the Net, such as law, regulation, or social pressure, restrain third parties that turn out to be untrustworthy, systems that turn out to protect one’s identity less well than promised, and so on. How satisfactory any of the nontechnical mechanisms may be depends on one’s expectations for the role of government (e.g., how paternalistic), the role of industry (e.g., how exploitative or how responsible), and the ability and willingness of individuals to become suitably informed and act in their own defense (in the case of privacy and security concerns) or responsibly (in the case of such concerns as taxation).[59]

There is a philosophical different between the technical and the legal approaches that have been discussed here. Technical mechanisms have the feature that their behavior is predictable *a priori*. One can examine the mechanism, convince oneself as to what it does, and then count on it to work as described. Legal mechanisms, on the other hand, often come into play after the fact. A party can go to court (a kind of third party), and as a result of a court order or injunction, achieve change; of course, the existence of a legal mechanism is generally associated with an expectation of deterrence.

For example, the nym server cited above addresses the problem of email anonymity through technical means. By the creative use of encryption, careful routing of data by the communicating application, and absence of logging, it becomes essentially impossible to determine after the fact who sent a message.[60] The result (beneficial in the eyes of the designers) is that one can use the nym server with the confidence that nobody, whether “good guy” or “bad guy” can later come in and force the revelation of the identity. The drawback is that “bad guys” might use cover of anonymity to do really bad things, bad enough to tip the balance of opinion toward response and away from protection of anonymity at all costs. Would society like a remedy in this case?

At a philosophical level, the debate itself represents an important part of finding the right balance. But for the moment, the Internet is a system where technology rather than law is the force most immediately shaping behavior, and until the legal environment matures, there is comparatively less option for remedy after the fact for actions in cyberspace than in real space.[61]

Some argue that law has limited value in influencing Internet-based conduct because the Internet is transborder, sources and destinations can be in unpredictable jurisdictions, and/or sources and destinations can be in jurisdictions with different bodies of law. This argument encourages those who would call for technical controls (which simply work the way they work, independent of jurisdiction and therefore of varying satisfaction to specific jurisdictional authorities), and those who argue for private, group-based self-regulation, where groups of users agree by choice on an approach (e.g., the use of PICS) to create a shared context in which they can function. Because of the limitations of private, group-based regulation, a variety of regulatory agencies is examining a variety of conditions relating to the conduct of business over the Internet and weighing options for intervention, in turn motivating new attempts at self-regulation that may or may not be effected or accepted. Meanwhile, legal solutions are being actively explored. [62]

Assessing where we are today

As noted in the introduction, many forces are pushing to change the Internet today: a greater call (from various voices) for stable and reliable operation, even though we can place less trust in the individual users of the network; new sorts of sophisticated applications driven by new visions of consumer-oriented experiences; the motivation of ISPs to develop into enclaves containing enhanced service to gain competitive advantage; the proliferation of third parties with a range of interests in what the users are actually doing; the proliferation of less sophisticated users for whom “innovation” is a mixed blessing; and new forms of computing and communication that call for new software structures. All of these forces have the consequences of increased complexity, of increased structure in the design of the Internet, and of a loss of control by the user. Whether one chooses to see these trends as a natural part of the growing up of the Internet or the fencing of the West, they are happening. It is not possible to turn back the clock to regain the circumstances of the early Internet: real changes underscore the real questions about the durability of the Internet’s design principles and assumptions.

The rise of the new players

Much of what is different about the Internet today can be traced to the new players that have entered the game over the last decade. The commercial phase of the Internet is really less than ten years old—NSFnet, the government-sponsored backbone that formed the Internet back in the 1980s, was only turned off in 1995. At that time, when the commercial ISPs began to proliferate, the number of players was very small, and their roles were fairly simple.

The world has become much more complex since that time. One trend is obvious: the changing role of the government in the Internet. The historic role of enabler is withering; comparatively speaking, government contributions to the design and operation of the Internet have shrunk.[63] At the same time, as more and more citizens have started to use the Internet and depend on it, government attention to the nature of Internet businesses and consumer issues has grown. This trend was easily predicted, even if viewed by some with regret. In fact the roles that the government is playing are consistent with government activities in other sectors and with the history of conventional telecommunications, including both telephony and broadcast media: antitrust vigilance, attempts to control consumer fraud, definition of a commercial code, taxation, and so on. There is little the government has done that represents a new role. In the telecommunications area the government has a special set of laws and a special agency, the Federal Communications Commission, to deal with presumed issues of natural monopoly and spectrum scarcity by translating law into regulation and attending to regulatory enforcement. In the United States, the government has largely refrained from bringing these tools to bear on the Internet, but the potential for doing so is widely recognized (not least because of scrutiny of mergers and acquisitions that bear on the development of the Internet) and has itself influenced the conduct of the players.

The wild card has been the development of the ISP. Its role is less clear and less predefined than that of the government, and it has evolved and become much more complex. Government recognized in the early 1990s that the private sector would build the National (eventually Global) Information Infrastructure, and the gold rush that ensued from commercializing the backbone made the ISP business resemble many others, with ISPs pursuing the most profitable means to define and carry out a business endeavor. Any action that an ISP undertakes to enhance its role beyond basic packet forwarding is not likely to be compatible with end to end thinking, since the ISP does not have control over the end-points. The ISP implements the core of the network, and the end-point software traditionally comes from other providers.[64] So the ISP is most likely to add services and restraints by modifying the part of the network that it controls. For example, some residential users find themselves blocked from running a Web or game server in their home.[65] Those services are restricted to commercial customers who pay a higher fee for their Internet access. From one perspective, such service stratification is only natural: it is in the nature of private enterprise to separate users into different tiers with different benefits and price them accordingly. Anyone who has flown at full fare while the person with the Saturday-night stay flies for a small fraction of the cost has understood value-based pricing. And yet some Internet observers have looked at such restrictions, when applied to Internet service, as a moral wrong. From that perspective, the Internet should be a facility across which the user should be able to do anything he wants, end to end. As a society, much less across all the societies of the world, we have not yet begun to resolve this tension.

Concerns about the final form of Internet service in an unconstrained commercial world are increased by industry consolidation, which raise concerns about adequate competition in local access (as marked by ATT's acquisition of TCI and MediaOne), and by mergers between Internet access providers and Internet content providers (marked by AOL's proposed acquisition of Time-Warner, including all its cable facilities).[66] A related issue is the "open access" debate, which concerns whether ISPs should be compelled to share their facilities. The concern is not just about choice in ISPs, but that if access to alternative ISPs is constrained or blocked, then users would be able to access some content only with difficulty, if at all. There is thus a presumed linkage between lack of choice in access to the Internet and a loss of the open, end to end nature of the Internet.[67]

As a broader base of consumers has attached to the Internet, they have sought out very different sorts of experiences. In the competitive world of dial-up Internet access, the company that holds the major share of U.S. consumers is America Online, or AOL. One can speculate about the sorts of experience that the consumer favors by looking at what AOL offers. The emphasis of AOL is less on open and equal access to any activity and destination (what the end to end arguments would call for), and more on packaged content (reinforced by the anticipated merger with Time Warner), predictable editorship, and control of unwelcome side-effects. Their growing subscribership attests to consumer valuation of the kind of service they offer and the comparative ease of use they provide. Those who call for one or another sort of Internet as a collective societal goal would at least do well to learn from the voice of the consumer as it has been heard so far.

New questions are arising about the legal treatment of ISPs. The rise of ISPs and transformations of historically regulated telephone companies, broadcasters, and more recently cable television providers have created new tensions between a broad goal of relaxing economic regulation—with the goals of promoting competition and such attendant consumer benefits as lower prices and product innovation—and concerns about the evolving structure and conduct of the emerging communications services leaders—factors shaping actual experience with prices and innovation. Although U.S. federal telecommunications regulators have eschewed "regulation of the Internet," topics being debated include whether the legal concept of common carriage that applies to telephone service providers should apply to ISPs.[68] Today's legislative and regulatory inquiries beg the question of whether the ISP business should continue to evolve on its own—whether the transformation of the Internet into public infrastructure calls for some kind of intervention.[69]

The institutional providers of Internet services—the corporations, schools and non-profit organizations that operate parts of the Internet—have also evolved a much more complex set of roles. Employees have found themselves fired for inappropriate use of the corporate attachment to the Internet, and employers have sometimes been much more restrictive than ISPs in the services they curtail and the rules they impose for acceptable use. The user of the Internet today cannot necessarily do as he pleases: he can do different things across different parts of the Internet, and perhaps at different times of the day.

Finally, one must never lose sight of the international nature of the Internet. As the Internet emerges and grows in other countries, which it is doing with great speed, the cultural differences in different places will be a major factor in the overall shape the Internet takes. In some countries, the ISP may be the same thing as the government, or the government may impose a set of operating rules on the ISPs that are very different from those we expect in the U.S.

The erosion of trust

A number of examples in this paper have illustrated that users who do not totally trust each other still desire to communicate. Of all the changes that are transforming the Internet, the loss of trust may be the most fundamental. The exact details of what service an ISP offers may change over time, and they can be reversed by consumer pressure or law. But the simple model of the early Internet—a group of mutually trusting users attached to a transparent network—is gone forever. To understand how the Internet is changing, we must have a more sophisticated consideration of trust and how it relates to other factors such as privacy, openness, and utility.

The spread of the Internet into more and more spheres of economic and social activity suggests growth in its use both among trusting and non-trusting parties. A result is growing individual interest in self-protection, something that may involve, actively or passively, third parties. Against this backdrop arise concerns of specific third parties to meet their own objectives, such as protection of assets, revenue streams, or some form of public safety. That is, trustworthiness motivates both self-protection (which may be end to end) and third-party intervention (which appears to challenge the end to end principles).

As trust erodes, both end-points and third parties may wish to interpose intermediate elements into a communication to achieve their objectives of verification and control. For intermediate elements interposed between communicating parties in real time, there is a tension between the need for devices to examine (at least parts of) the data stream and the growing tendency for users and their software to encrypt communication streams to ensure data integrity and control unwanted disclosure. If a stream is encrypted, it cannot be examined; if it is signed, it cannot be changed. Historically, encryption for integrity protection has been accepted more easily by authorities concerned about encryption than encryption for confidentiality, but that may be too glib an assumption in a world with pervasive encryption, where individuals may encounter circumstances when encryption is not an unmitigated good. For example, in the real world, one shows caution about a private meeting with a party that one does not trust. One seeks a meeting in a public place, or with other parties listening, and so on. Having an encrypted conversation with a stranger may be like meeting that person in a dark alley. Whatever happens, there are no witnesses. Communication in the clear could allow interposed network elements to process the stream, which could be central to the safety and security of the interaction. This example of a case where an individual might choose to trade off privacy for other values illustrates the proposition that choices and tradeoffs among privacy, security, and other factors are likely to become more complicated.

At the same time, there are many transactions that the collection of end-points may view as private, even though there is not total trust among them. In an online purchase, details such as the price or the credit card number might deserve protection from outside observation, but the fact of the purchase might be a matter of record, to provide a basis for recourse if the other party misbehaves. Such situations may argue for selective use of encryption—not the total encryption of the data stream at the IP level (as in the IPsec proposal), but applied selectively, for example by the browser to different parts of a message. The use of IPsec would most naturally apply to communication among parties with the highest level of trust, since this scheme protects the maximum amount of information from observation.

The use of trusted third parties in the network raises the difficulty of how one can know that third parties are actually trustworthy, or that the end-points are talking to the third party they think they are. What happens if a malicious “imitation” third party manages to insert itself in place of a trusted agent? Today, Web sites attempt to snare the unwary using names similar to respected ones. How can the users of the Internet be confident that sites that are physically remote, and only apparent through their network behavior, are actually what they claim, actually worthy of trust?[70]

Rights and responsibilities

The rise of legal activity reflects the rise of debates that center on the relative power (or relative rights, or relative responsibility) that devolves to the end users as individuals and to the network as an agent of the common good (e.g., the state, the group of users served by a given network). Some of these debates are rooted in law of a country or state, some in value systems and ideology. The First Amendment to the U.S. Constitution speaks to a positive valuation of free speech; other countries have different normative and legal traditions. Similarly, societies will differ in how they define accountability and in how they strike a balance between anonymity and accountability. Given differing national contexts, different geographically defined regions of the network may be managed to achieve differing balances of power,[71] just as different organizations impose different policies on the users of their networks. Local control may be imperfect, but it does not have to be perfect to shape the local experience. But if the Internet is to work as an internetwork, there are some limits on just how different the different regions can be.

The end to end design of the Internet gives the user considerable power in determining what applications he chooses to use. This power raises the possibility of an “arms race” between users and those who wish to control them. That potential should be a sobering thought, because it would have quite destructive side-effects. The cryptography policy debate held that if, for example, controls were put in the network that attempted to intercept and read private communications between parties, the response from the users could easily be to encrypt their private communication. The response to that would either be to outlaw the use of encryption, to promote government-accessible keys, or to block the transmission of any message that cannot be recognized, which might in turn lead to messages hidden inside other messages—steganography. It would seem that an attempt to regulate private communication, if it were actually feasible to implement (such controls seem to be getting harder), would result in a great loss of privacy and privilege for the affected individuals.[72] These sorts of controls also serve to block the deployment of any new application, and stifle innovation and creativity. Consider what the Internet might look like today if one had to get a license to deploy a new application. This sort of escalation is not desirable.

Perhaps the most critical tension between rights and responsibilities is one that emerges from the erosion of trust—it is the balance between anonymity and accountability. The end to end arguments, by their nature, suggest that end-points can communicate as they please, without constraint from the network. This implies, on the one hand, a certain need for accountability, in case these unconstrained activities turn out to have caused harm. Any system, whether technical or societal, requires protection from irresponsible and harmful actions. The end to end arguments do not imply guard rails to keep users on the road. On the other hand, there has been a call for the right of anonymous action, and some sorts of anonymous actions (such as political speech in the United States) are a protected right. Certainly privacy, if not absolute anonymity, is a much-respected objective in many societies. So how can the desire for privacy and anonymity be balanced against the need for accountability, given the freedom of action that the end to end arguments imply? This will be a critical issue in the coming decade.

A practical issue in moving forward is the enforceability of a policy. Some kinds of communications, and some kinds of parties, are more tractable when it comes to implementing controls (or behavior that obviates a need for controls in the eyes of those with concerns). For example, there is a distinction that often recurs: the separation between private and public communication. Today, the Internet places few limits on what two consenting end-nodes do in communicating across the network. They can send encrypted messages, design a whole new application, and so on. This is consistent with the simple articulation of the end to end arguments. Such communication is *private*. In contrast, *public* communication, or communication *to the public,* has different technical and social characteristics.

- In order to reach the public, one must advertise.
- In order to reach the public, one must use well-known protocols and standards that the public has available.
- In order to reach the public, one must reveal one’s content. There is no such thing as a public secret.
- In order to reach the public, one must accept that one may come under the scrutiny of the authorities.

These factors make public communication much easier to control than private communication, especially where public communication is commercial speech (where, to a limited degree, at least in the United States, more rules can be applied than to noncommercial speech). In the case of labels on information that is otherwise encrypted, the authorities may not be able to verify that every label is proper. But authorities can check whether the sender is computing proper labels by becoming a subscriber to the service, seeing if the information sent is properly labeled.[73]

Another pattern of communication that supports enforcement is between an individual and a recognized institution. In many cases, one end of a transfer or the other may be easier to hold accountable, either because it is in a particular jurisdiction, or because it is a different class of institution. For example, it may be easier to identify and impose requirements on corporations and other businesses, compared to individuals. Thus, in a transaction between a customer and a bank, it may be easier to impose enforceable regulation on the bank than the client. Banks are enduring institutions, already subjected to much regulation and auditing, while the individual customer is less constrained. This can create a situation in which the bank becomes part of the enforcement scheme. Similarly, providers of content, if they are intending to provide that content to the public, are of necessity more identifiable in the market than the individual customer, and that makes them visible to enforcement agencies as well as to their desired customers. Even if one can not check their correct behavior on every transfer from a content provider, the legal authorities can perform a spot-check, perhaps by becoming a customer. If the penalties for non-compliance are substantial, there may be no need to verify the accuracy of every transfer to achieve reasonable compliance.[74] Recognition and exploitation of these differing roles for institutions and for individuals may enhance the viability of end-located applications and the end to end approach in general.

Conclusions

The most important benefit of the end to end arguments is that they preserve the flexibility, generality, and openness of the Internet. They permit the introduction of new applications; they thus foster innovation, with the social and economic benefits that follow. Movement to put more functions inside the network jeopardizes that generality and flexibility as well as historic patterns of innovation. A new principle evident already is that elements that implement functions that are invisible or hostile to the end to end application, in general, have to be “in” the network, because the application cannot be expected to include that intermediate element voluntarily.

Multiple forces seem to promote change within the Internet that may be inconsistent with the end to end arguments. While there has been concern expressed in some quarters about the increasing involvement of governments, the ISP may present a greater challenge to the traditional structure of the Internet. The ISPs implement the core of the network, and any enhancement or restriction that the ISP implements is likely to appear as new mechanism in the core of the network. As gateways to their customers they are an inherent focal point for others interested in what their customers do, too.

The changing nature of the user base is pushing the Internet in new directions, contributing to both ISP and government efforts. At issue is the amount of end-point software owned and operated, if not understood, by consumers and therefore the capacity of the Internet system in the large to continue to support an end to end philosophy. While the original Internet user was technical and benefited from the flexibility and empowerment of the end to end approach, today’s consumer approaches the Internet and systems like other consumer electronics and services. Low prices and ease of use are becoming more important than ever, suggesting growing appeal of bundled and managed offerings over do it yourself technology. Less work by consumers may imply less control over what they can do on the Internet and who can observe what they do; the incipient controversy over on-line privacy, however, suggests that there are limits to what many consumers will cede for various reasons.

Of all the changes that are transforming the Internet, the loss of trust may be the most fundamental. The simple model of the early Internet—a group of mutually trusting users attached to a transparent network—is gone forever. A motto for tomorrow may well be “global communication with local trust.” Trust issues arise at multiple layers: within Internet-access (e.g., browser) and application software (some of which may trigger Internet access), within activities that access content or effect transactions out at remote sites, within communications of various kinds with strangers, and within the context of access networks—operated by ISPs, employers, and so on—whose operators seek to attend to their own objectives while permitting others to use their networks. Growing concern about trust puts pressure on the traditional Internet support for anonymity. The end to end arguments, by their nature, suggest that end-points can communicate as they please, without constraint from the network, and at least in many Western cultures anonymity is valued in many contexts. Growth in societal use and dependence on the Internet, however, induces calls for accountability (itself varied in meaning), creating pressures to constrain what can happen at end-points or to track behavior, potentially from within the network. One step that can support trust in some contexts is to provide systematic labeling of content. As ongoing experiments suggest, labeling may assist in protection of privacy, avoidance of objectionable material, and anonymity while preserving end to end communications, but they still pose significant technical and legal challenges.

More complex application requirements are leading to the design of applications that depend on trusted third parties to mediate between end users, breaking heretofore straightforward end to end communications into series of component end to end communications. While this approach will help users that do not totally trust each other to have trustworthy interactions, it adds its own trust problems: how one can know that third parties themselves are actually trustworthy, or that the end-points are talking to the third party they think they are? It doesn’t take too many of these options to realize that resolving Internet trust problems will involve more than technology, and the proliferation of inquiries and programmatic actions by governments plus a variety of legal actions combine to impinge on the Internet and its users.

It may well be that certain kinds of innovation would be stifled if the open and transparent nature of the Internet were to erode. Today there is no evidence that innovation has been stifled overall. The level of investment in new dot-com companies and the range of new offerings for consumers, ranging from e-commerce to online music, all attest to the health of the evolving Internet. But the nature of innovation may have changed. It is no longer the single creative person in the garage, but the startup with tens of millions of dollars in backing that is doing the innovation. And it may be that the end to end arguments favor the small innovator, while the more complex model of today, with content servers and ISP controls on what services can and cannot be used in what ways, are a barrier to that small innovator, but not to the well-funded innovator who can deal with all these issues as part of launching a new service. So the trend for tomorrow may not be the simple one of slowed innovation, but the more subtle one of innovation by larger players backed with more money.

Perhaps the most insidious threat to the end to end arguments, and thus to flexibility, is that commercial investment will go elsewhere, in support of short-term opportunities better met by solutions that are not end to end, but based on application-specific servers and services “inside” the network. Content mirroring, which positions copies of content near the consumer for rapid, high performance delivery, facilitates the delivery of specific material, but only material that has been mirrored. Increasing dependence on content replication might reduce investment in general-purpose upgrades to Internet capacity. It is possible that we will see, not a sudden change in the spirit of the Internet, but a slow ossification of the form and function. In time some new network will appear, perhaps as an overlay on the Internet, which attempts to re-introduce a context for unfettered innovation. The Internet, like the telephone system before it, could become the infrastructure for the system that comes after it.

We have painted two pictures of the constraints that technology imposes on the future Internet. One is that technological solutions are fixed and rigid. They implement some given function, and do so uniformly independent of local needs and requirements. They create a black-and-white outcome in the choice of alternatives. Either an anonymizing service exists, or it does not. On the other hand, we observe in practice that there is a continuing tussle between those who would impose controls and those who would evade them. There is a tussle between spammers and those who would control them, between merchants who need to know who the buyers are and buyers who use untraceable e-mail addresses, and between those who want to limit access to certain content and those who try to reach it. This pattern suggests that the balance of power among the players is not a winner-take-all outcome, but an evolving balance. It suggests that the outcome is not fixed by specific technical alternatives, but the interplay of the many features and attributes of this very complex system. And it suggests that it is premature to predict the final form. What we can do now is push in ways that tend toward certain outcomes. We argue that the open, general nature of the Net, which derived from the end to end arguments, is a valuable characteristic that encourages innovation, and this flexibility should be preserved.

---

1 Clark's research is supported by the Defense Advanced Research Projects Agency under contract N6601-98-8903, and by the industrial partners of the M.I.T. Internet Telecomms Convergence Consortium. Blumenthal is an employee of the complex derived from the National Academy of Sciences, and when this paper was framed in 1998 was also an employee of M.I.T.. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policy or endorsements, either expressed or implied, of DARPA, the U.S. Government, or the National Academies.

2 See Saltzer, J., Reed, D., and Clark, D.D.. 1984. "End-to-End Arguments in System Design." *ACM Transactions on Computer Systems*, Vol. 2, No. 4, November, pp. 277-288.

3 See Computer Science and Telecommunications Board. 1999. *Trust in Cyberspace*, National Academy Press.

4 For one view of spam and its control, see D. Dorn, 1998, "Postage due on junk e-mail—Spam costs Internet millions every month" *Internet Week*, May 4, 1998; at http://www.techweb.com/se/directlink.cgi?INW19980504S0003. For a summary of legislative approaches to control of spam, see Ouellette, Tim. 1999. "Technology Quick Study: Spam." *Computerworld*, April 5, p.70. The Mail Abuse Prevention System (MAPS.LLC), provides tools for third parties (ISPs) to filter and control spam. Their charter states that their approach to control of spam is "educating and encouraging ISP's to enforce strong terms and conditions prohibiting their customers from engaging in abusive e-mail practices." See http://www.mail-abuse.org/.

5 There has been a great deal of work over the last decade to define what are called Quality of Service mechanisms for the Internet. See Braden, R, D. Clark and S. Shenker. 1994. *Integrated services in the Internet Architecture: an overview*. RFC 1633, IETF, and Carlson, M., et al. 1998. *An Architecture for Differentiated Services*. RFC 2475, IETF. The progress of this work is reported at http://www.ietf.org/html.charters/intserv-charter.html and http://www.ietf.org/html.charters/diffserv-charter.html.

6 See Larson, Gary and Jeffrey Chester. 1999. *Song of the Open Road: Building a Broadband Network for the 21st Century*. The Center for Media Education Section IV, p 6. Available at http://www.cme.org/broadband/openroad.pdf.

7 We also discuss other kinds of third parties, whose services may be sought out by the communicating end-points or whose actions are otherwise tolerated by them. There is growing potential for both kinds of third parties, but this section focuses on the imposition of unwelcome third parties.

8 This trend is signaled by the rise of the Application Service Provider, or ASP, as a part of the landscape.

9 A common method for constructing "configuration free," or "plug and play," or "works out of the box" devices is to assume that some other element takes on the role of controlling setup and configuration. Of course, centralization raises other issues, such as a common point of vulnerability, and the proper balance is not yet clear between centralization and distribution of security function for consumer networking.

10 For example, see: Saltzer, Jerome H. 1999. *"Open Access" is just the tip of the iceberg.* October 22, available at http://web.mit.edu/Saltzer/www/publications/openaccess.html. and Lemley, Mark A. and Lawrence Lessig. 1999. Filing before the Federal Communications Commission, (In the Matter of Application for Consent to the Transfer of Control of Licenses

MediaOne Group, Inc. to AT&T Corp. CS Docket No. 99-251). Available at http://cyber.law.harvard.edu/works/lessig/MB.html. Lessig's work can be seen in overview at http://cyber.law.harvard.edu. For a lightweight example that speaks directly to end to end, see: Lessig, Lawrence. 1999. "It's the Architecture, Mr. Chairman."

11 The Electronic Signatures in Global and National Commerce Act is an indicator of the broadening recognition of a need for tools to support network-mediated transactions, although observers note that it raises its own questions about how to do so—resolving the technology and policy issues will take more work.

12 Chaum, David. 1992. "Achieving Electronic Privacy." *Scientific American*. August. pp. 96-101.

13 It may seem that this attention to protection of identity, especially as it manifests in low-level information such as addresses, is exaggerated. The telephone system provides an illustration of how attention to identity has grown and added complexity to communications. For most of the history of the telephone system, the called telephone (and thus the person answering the phone) had no idea what the number of the caller was. Then the "caller ID" feature was invented, to show the caller's number to the called party. This very shortly led to a demand for a way to prevent this information from being passed across the telephone network. Adding this capability, which re-instituted caller anonymity at the level of the phone number, led in turn to demand for the feature that a receiver could refuse to receive a call from a person who refused to reveal his phone number. Additional issues have arisen about the treatment of phone numbers used by people who have paid for "unlisted" numbers, which appears to vary by telephone service provider and state regulatory decision. Given the emergence of this rather complex balance of power in conventional telephony, there is no reason to think that users of the Internet will eventually demand any less. Even if the identity of the individual user is not revealed, this low level information can be used to construct profiles of aggregate behavior, as in Amazon's summer 1999 publicity about book-buying patterns of employees of large organizations based on e-mail addresses.See Amazon.com. 1999. "Amazon.com Introduces 'Purchase Circles [TM],' Featuring Thousands of Bestseller Lists for Hometowns, Workplaces, Universities, and More." Press Release, Seattle, August 20, available at www.amazon.com; McCullagh, Declan. 1999. "Big Brother, Big 'Fun' at Amazon." *Wired*, August 25, available at www.wired.com/news/news/business/story/21417.html; Reuters. 1999. "Amazon modifies purchase data policy." *Zdnet*, August 27, available at www.zdnet.com/filters/printerfriendly/0,6061,2322310-2,00.html; and Amazon.com. 1999 "Amazon.com Modifies "Purchase Circles[TM]" Feature." Press Release, Seattle, August 26, available at www.amazon.com.

14 An example of this give and take is the popularity of e-mail accounts from a provider such as Hotmail that does not require the user to prove who he really is (as would be required where a financial account is established). This permits the user to send messages with relative anonymity. As a result of this, some online merchants will not accept orders from users who use Hotmail accounts.

15 Cookies may be part of a larger class of monitoring software. See, for example, O'Harrow, Jr., Robert. 1999. "Fearing a Plague of 'Web Bugs': Invisible Fact-Gathering Code Raises Privacy Concerns." *Washington Post*, November 13, E1, E8.

16 See O'Harrow, R and E. Corcoran. 1999. "Intel Drops Plans for ID Numbers," *Washington Post*, January 26. http://www.washingtonpost.com/wp-srv/washtech/daily/jan99/intel26.htm. Intel backed away from use of the ID as an identifier in e-commerce transactions under consumer pressure . See http://www.bigbrotherinside.com/.

17 Microsoft implemented a scheme to tag all documents produced using Office 97 with a unique ID derived from the network address of the machine. In response to public criticism, they made it possible to disable this feature. They also discontinued the reporting of the hardware unique ID of each machine during online registration of Windows 98. See http://www.microsoft.com/presspass/features/1999/03-08custletter2.htm.

18 See Cha, Ariana Eunjung. 2000. "Your PC Is Watching: Programs That Send Personal Data Becoming Routine." *The Washington Post*, July 14, A1, A12-13.

19 See Computer Science and Telecommunications Board. 2000. *The Digital Dilemma: Intellectual Property in the Information Age*, National Academy Press.

20 D'Antoni, H. 2000. "Web Surfers Beware: Someone's Watching." *InformationWeek Online* , February 7, http://www.informationweek.com/bizint/biz772/72bzweb.htm. Examples of currently available software include SurfWatch, at http://www1.surfwatch.com/products/swwork.html, and Internet Resource Manager, at http://www.sequeltech.com/.

21 The rash of denial of service attacks on major Web sites in early 2000 illustrates the magnitude of this problem.

22 Moss, Michael. 1999. "Inside the game of E-Mail Hijacking." *The Wall Street Journal*, November 9, B1, B4. "Already, the Internet is awash in Web sites that trick people into clicking on by using addresses that vary only slightly from the sites being mimicked: an extra letter here, a dropped hyphen there. Now, in near secrecy, some of these same look-alike Web sites are grabbing e-mail as well."

23 A series of publicized problems affecting Microsoft's Internet Explorer, and the generation of associated software fixes, is documented on the Microsoft security site: http://www.microsoft.com/windows/ie/security/default.asp. A similar list of issues for Netscape Navigator can be found at http://home.netscape.com/security/notes/.

24 Jerome Saltzer, 1998. Personal communication, Nov 11.

25 As opposed to taxation of the use of the Internet per se, like taxation of telephone service. This discussion does not address the merits of taxation; it proceeds from the recognition of (multiple) efforts to implement it.

26 For example, independent of technology, income tax compliance is promoted by the practice—and risk—of audits.

27 Practically, many pornography sites today use the combination of possession of a credit card and a self-affirmation of age as an acceptable assurance of adulthood—although some minors have credit cards. Indicating adulthood has different ramifications from indicating minority, as Lessig has noted; the intent here is to contrast identification of content and users.

28 There are other purposes for which a control point "in" the net might be imposed, to achieve a supposedly more robust solution than an end-point implementation can provide. These include facilitating eavesdropping/wiretap, collection of taxes and fees associated with transactions using the network, and so on. One question now being discussed in the Internet Engineering Task Force (IETF) is how, if at all, Internet protocols should be modified to support Communications Assistance for Law Enforcement Act of 1995 (CALEA) wiretap regulations. See Clausing, Jeri. 1999. "Internet Engineers Reject Wiretap Proposal." *The New York Times*, November 11, B10. The current sentiment in the design community is that this is not an appropriate goal for the IETF. However, there appears to be some interest from equipment vendors in conforming to CALEA, given interest expressed by their customers, so the outcome of this discussion remains unclear.

29 It is possible that the introduction of the new Internet address space, as part of the next generation Internet protocol called IPv6, with its much larger set of addresses, will alleviate the need for NAT devices. There is much current debate as to whether NAT devices are a temporary fix, or now a permanent part of the Internet.

30 As this paper was being completed, news broke about the FBI's "Carnivore" system, characterized as an "Internet wiretapping system" that is deployed at an ISP's premises. See King, Neil, Jr., and Ted Bridis. 2000. "FBI's Wiretaps To Scan E-Mail Spark Concern." *The Wall Street Journal*, July 11, A3, A6. Also, note that users who move from place to place and dial in to different phone numbers do not use the same physical link for successive access, but since they have to authenticate themselves to the ISP to complete the connection, the ISP knows who is dialing, and could institute logging accordingly.

31 Similarly, if an organization has any requirement imposed on it to control the behavior of its users, it will be at the point of egress that the control can best be imposed.

32 Of course, this sort of control is not perfect. It is possible for a creative user to purchase a number of ISP accounts and move from one to another in an unpredictable way. This is what is happening today in the battle between spammers and those who would control them, another example of the dynamic tussle between control and avoidance.

33 California Assembly Bill1676, enacted 1998.

34 For a detailed discussion of labels on content and on users, see Lessig, Lawrence and Paul Resnick (1999). "Zoning Speech on the Internet: A Legal and Technical Model." *Michigan Law Review* 98(2): 395-431.

35 This is a critical issue for the viability of industry self-regulation. That topic, given the looming prospect of government regulation, is the subject of much debate. Major industry players and scholars, for example, participated in a 1999 international conference organized by the Bertelsmann Foundation, which cast labeling approaches as user-empowering and urged government support for private filtering based on labeling. See Bertelsmann Foundation. 1999. *Self-regulation of Internet Content*. Gutersloh, Germany, September, available at http://www.stiftung.bertelsmann.de/internetcontent/english/content/c2340.htm.

36 See, for example: U.S. Federal Trade Commission. 1998. *Advertising and Marketing on the Internet: Rules of the Road*. Washington, DC, August, available at www.ftc.gov.

37 The PICS web site maintained by the World Wide Web Consortium is http://www.w3.org/pics.

38 There are a number of Web proxy servers that implement PICS filtering. See http://www.n2h2.com/pics/proxy_servers.html.

39 For a discussion of concerns aroused by PICS, see http://rene.efa.org.au/liberty/label.html. For a response to such concerns by one of the PICS developers and proponents, see Resnick, Paul, ed. 1999. "PICS, Censorship, & Intellectual Freedom FAQ." Available at www.w3.org/PIC/PICS-FAQ-980126.HTML.

40 The Metatdata web site maintained by the World Wide Web Consortium is http://www.w3.org/Metadata/.

41 For example, there have been lawsuits attempting to prevent the use of a trademark in the metadata field of a page not associated with the holder of the mark. A summary of some lawsuits related to trademarks in metadata can be found at http://www.searchenginewatch.com/resources/metasuits.html.

42 Examples of anonymizing browser services can be found at http://www.anonymizer.com, http://www.idzap.net/, http://www.rewebber.com/, http://www.keepitsecret.com/, http://www.confidentialonline.com/home.html, and http://www.websperts.net/About_Us/Privacy/clandestination.shtml. The last of these offers a service in which the anonymous intermediate is located in a foreign country to avoid the reach of the U.S. legal system. The quality of some of these services is questioned in Oakes, Chris, 1999, "Anonymous Web Surfing? Uh-Uh," *Wired News*, Apr. 13, http://www.wired.com/news/technology/0,1282,19091,00.html.

43 For one example of a system that tries to provide protection from traffic analysis, see Goldschlag, David M., Michael G. Reed, and Paul F. Syverson. 1999. "Onion Routing for Anonymous and Private Internet Connections." *Communications of the ACM*, vol. 42, num. 2, February. For a complete bibliography and discussion, see http://onion-router.nrl.navy.mil/.

44 Mazières, David and M. Frans Kaashoek. 1998. "The design, implementation and operation of an email pseudonym server." *Proceedings of the 5th ACM Conference on Computer and Communications Security* (CCS-5). San Francisco, California, November, pages 27-36.

45 The outgoing message is prefaced with a sequence of addresses, each specifying a relay point. Each address is encrypted using the public key of the prior hop, so that the relay point, and only the relay point, can decrypt the address of the next hop the message should take, using its matching private key. Each relay point delays the message for an unpredictable time, so that it is hard to correlate an incoming and an outgoing message. If enough hops are used, it becomes almost impossible to trace the path from destination back to the source.

46 For a review of tools currently available to filter spam in mail servers, see http://spam.abuse.net/tools/mailblock.html.

47 More complex replication/hosting schemes for controlled staging of content provide features to remedy these limitations, in return for which the content provider must usually pay a fee to the service.

48 This is a topic that has been receiving more analysis in different contexts. For a legal assessment, see, for example, Froomkin, A. Michael. 1996. "The Essential role of Trusted Third Parties in Electronic Commerce," *Oregon Law Review* 75:29, available at www.law.miami.edu/~froomkin/articles/trustedno.htm.

49 For example, see the mutual commitment protocol in Jianying Zhou, Dieter Gollmann. 1996 "A Fair Non-repudiation Protocol." *Proceedings of the 1996 Symposium on Security and Privacy*, Oakland, May 6-8.

50 A notary is "[a] responsible person appointed by state government to witness the signing of important documents and administer oaths." See National Notary Association. 1997. "What is a Notary Public?" Chatsworth, CA, at http://www.nationalnotary.org/actionprograms/WhatisNotaryPublic.pdf. Recognition of this role has led to the investigation of a "cyber-notary" as a useful agent within the Internet This has been a topic of study by the American Bar Association, but there does not appear to be an active interest at this time.

51 There is a partial analogy with payment by check, where the bank balance is normally not verified at the moment of purchase. However, the taker of the check may demand other forms of identification, which can assist in imposing a fee for a bad check. If a certificate has been invalidated, the recipient cannot even count on knowing who the other party in the transaction actually is. So there may be fewer options for later recourse.

52 We emphasize the broader choice of mechanism out of the recognition that technologists often prefer technical solutions. The Internet philosophy acknowledged early in the paper argues for the superiority of technology over other kinds of mechanisms. See, for example, Goldberg, Ian, David Wagner, and Eric Brewer. 1997. "Privacy-enhancing technologies for the Internet," available at www.cs.berkeley.edu/~daw/privacy-compcon97-222/privacy-html.html. Those authors observe that "[t]he cyperpunks credo can be roughly paraphrased as 'privacy through technology, not through legislation.' If we can guarantee privacy protection through the laws of mathematics rather than the laws of men and whims of bureaucrats, then we will have made an important contribution to society. It is this vision which guides and motivates our approach to Internet privacy."

53 There is no technical verification that this number is indeed sent (fax is, like the Internet, very much an end to end design), but the presumption is that the law can be used to keep the level of unwanted faxes to an acceptable level. Note also that this law, which had the goal of controlling receipt of unwanted material, outlaws "anonymous faxes," in contrast to telephone calls, where one can prevent the caller's phone number from being passed to the called party.

54 This trend was emphasized by the mid-1999 establishment, by executive order, of a federal task force concerned with illegal conduct on the Internet. President's Working Group on Unlawful Conduct on the Internet. 2000. *The Electronic Frontier: The Challenge of Unlawful Conduct Involving the Use of the Internet*. March. Available at: http://www.usdoj.gov/criminal/cybercrime/unlawful.htm.

55 The authors recognize that today on the Internet various labels are associated with voluntary schemes for content rating, etc.; illustrations of the complementarity of law or regulation come, at present, from other domains. Note, however, that the Bertelsmann Foundation conference summary cited above specifically cast law enforcement as a complement to voluntary labeling. It observed: "Law enforcement is the basic mechanism employed within any country to prevent, detect, investigate and prosecute illegal and harmful content on the Internet. This state reaction is essential for various reasons: It guarantees the state monopoly on power and public order, it is democratically legitimized and directly enforceable and it secures justice, equity and legal certainty. However, a mere system of legal regulation armed with law enforcement would be ineffective because of the technical, fast-changing and global nature of the Internet. In a coordinated approach, self-regulatory mechanisms have to be combined with law enforcement as a necessary backup." (p.45).

56 U.S. Federal Communications Commission, "V-Chip Homepage," available at http://www.fcc.gov/vchip/.

57 Information on Amazon.Com was cited above. On RealNetworks, see: Clark, Don. 1999. "RealNetworks Will Issue Software Patch To Block Its Program's Spying on Users." *The Wall Street Journal*, November 2, B8. That article explains, "Unbeknownst to users, the [Real-Jukebox] software regularly transmitted information over the Internet to the company,

including what CDs users played and how many songs were loaded on their disk drives." DoubleClick presented a broader privacy challenge because it tracked consumer movement across sites and products; the controversy it caused precipitated broad reactions, including government investigation due to a complaint made to the Federal Trade Commission. See: Tedeschi, Bob. 2000. "Critics Press Legal Assault on Tracking of Web Users." *The New York Times*, February 7, C1, C10.

58 Simpson, Glenn R. 2000. "E-Commerce Firms Start to Rethink Opposition To Privacy Regulation as Abuses, Anger Rise." *The Wall Street Journal*, January 6, A24.

59 What individuals can do for themselves, and what industry does, depends, of course, on incentives, which are a part of the nontechnical mechanism picture. Recent controversy surrounding the development of UCITA illustrates differing expectations and interpretations of who incurs what costs and benefits. An issue with these evolving frameworks is the reality that consumers, in particular, and businesses often prefer to avoid the costs of litigation.

60 The operators of the server are happy to provide what information they have in response to any court order, but the system was carefully designed to make this information useless.

61 This tension between technology, law, and other influences on behavior is at the heart of the much-discussed writing of Lawrence Lessig on the role of "code" (loosely, technology). See his 1999 book, *Code and Other Laws of Cyberspace,* Basic Books, New York. Critical responses to *Code…* note that technology is malleable rather than constant—a premise for this paper—and so are government and industry interests and motives. See, for example, Mann, Charles C. 1999. "The Unacknowledged Legislators of the Digital world." *Atlantic Unbound*, December 15, available at www.theatlantic.com/unbound/digicult/dc991215.htm.

62 What is known as "conflict of laws" provides a set of principles and models for addressing legal problems that span at least two jurisdictions. Resolving such problems is hard in the context of real space, and cyberspace adds additional challenges, but progress under the conflict of laws rubric illuminates approaches that include private agreements on which laws will prevail under which circumstances, international harmonization (difficult and slow but already in progress), and indirect regulation, which targets the local effects (e.g., behavior of people and equipment) of extraterritorial activity. For an overview, see Goldsmith, Jack L. 1998. "Against Cyberanarchy." *The University of Chicago Law Review*, 65:4, Fall, pp. 1199-1250. Among other things, Goldsmith explains that: "Cyberspace presents two related choice-of-law problems. The first is the problem of complexity. This is the problem of how to choose a single governing law for cyberspace activity that has multiple jurisdictional contacts. The second problem concerns situs. This is the problem of how to choose a governing law when the locus of activity cannot easily be pinpointed in geographical space." (p.1234) Case law shows that these issues are being worked out (or at least worked on). See, for example: Fusco, Patricia. 1999. "Judge rules ISP, Server Location May Determine Jurisdiction." ISP-Planet, June 11, available at www.isp-planet.com/politics/061199jurisdiction.html; and Kaplan, Carl S. 1999. "Judge in Gambling Case Takes On Sticky Issue of Jurisdiction." *The New York Times*, August 13, p.B10. The latter addressed the interplay of state law with federal law, which proscribes gambling via the Wire Act (18 USC 1084) and the Travel Act (18 USC 1952) and the Interstate Transportation of Wagering Paraphernalia Act (18 USC 1953). Some of these issues have been attacked by the American Bar Association's Internet Jurisdiction Project; see http://www.kentlaw.edu/cyberlaw/.

63 See Computer Science and Telecommunications Board. 1994. *Realizing the Information Future: The Internet and Beyond*, National Academy Press, and Computer Science and Telecommunications Board. 1999. *Funding a Revolution: Government Support for Computing Research*, National Academy Press.

64 Large ISPs such as AOL have attempted to attain control over the end nodes by distributing their own browser, which they encourage or require the user to employ. This approach has proved successful to some extent. In the future, we can expect to see ISP interest in extending their control over the end-point to the extend possible, for example by means of added function in Internet set top boxes and other devices they install in the home.

65 For example, see the Appropriate Use Policy of Excite@Home, at http://www.home.com/aup/, which specifically prohibits the operation of servers over their residential Internet service.

66 For an assessment of possible outcomes, see Saltzer, Jerome. 1999. "Open Access" is Just the Tip of the Iceberg," essay prepared for the Newton, MA Cable Commission, October 22, at http://mit.edu/Saltzer/www/publications/openaccess.html. After succinctly commenting on a number of possible outcomes that he finds undesirable, Saltzer notes that the most dire possible outcome of today's open access tussle, without open access and stifled competition and innovation, "is looking increasingly unlikely, as customers and cable competitors alike begin to understand better why the Internet works the way it does and the implications of some of the emerging practices."

67 See material cited in end-note 10 above. Note also the concerns raised under the rubric of "peering." See, for example, Caruso, Denise. 2000. "Digital Commerce: The Internet relies on networks' passing data to one another. But what happens if one of them refuses?" *The New York Times*, February 14, p.C4.

68 Common carriage implies certain rights and certain responsibilities, such as the provider's obligation to serve all comers while being protected from liability if those subscribers use the network for unacceptable purposes. The fact that the Internet has been designed such that (by the end to end arguments) ISPs cannot easily control the content sent over their networks and the fact that ISPs appear to serve all comers have caused some to suggest that ISPs be treated as common carriers; the suggestion also arises from those who perceive a greater ability of ISPs to control content than their nominal business and technology would suggest.

69 The late 1990s development of concern about "critical infrastructure" intensifies the attention and concern associated with growing reliance on the Internet, with explorations by the government and some industry leaders of new programs and mechanisms for monitoring use or "abuse" of the Internet and for increasing its robustness against malicious or accidental disruption. See Blumenthal, Marjory S. 1999. 1999. "Reliable and Trustworthy: The Challenge of Cyber-Infrastructure Protection at the Edge of the Millennium, " *iMP Magazine*, September, http://www.cisp.org/imp/september_99/09_99blumenthal.htm.

70 The popular fictional character Harry Potter receives some advice that might apply equally to his world and the Internet: "Never trust anything that can think for itself if you can't see where it keeps its brain." Rowling, J.K. 1998. *Harry Potter and the Chamber of Secrets*. Bloomsbury Publishing, London, p. 242.

71 Pomfret, John. 2000. "China Puts Clamps on Internet; Communists Seek Information Curb," *The Washington Post*, January 27.

72 See Computer Science and Telecommunications Board. 1996. *Cryptography's Role in Securing the Information Society*. National Academy Press.

73 Already today regulatory agencies (e.g., the Federal Trade Commission) are doing spot-checks of actual Web sites.

74 This approach is somewhat similar to the practice in some parts of the world of not always checking that passengers on public transit have the proper ticket in hand. Instead, there are roving inspectors that perform spot-checks. If the fine for failing to have the right ticket is high enough, this scheme can achieve reasonable compliance.