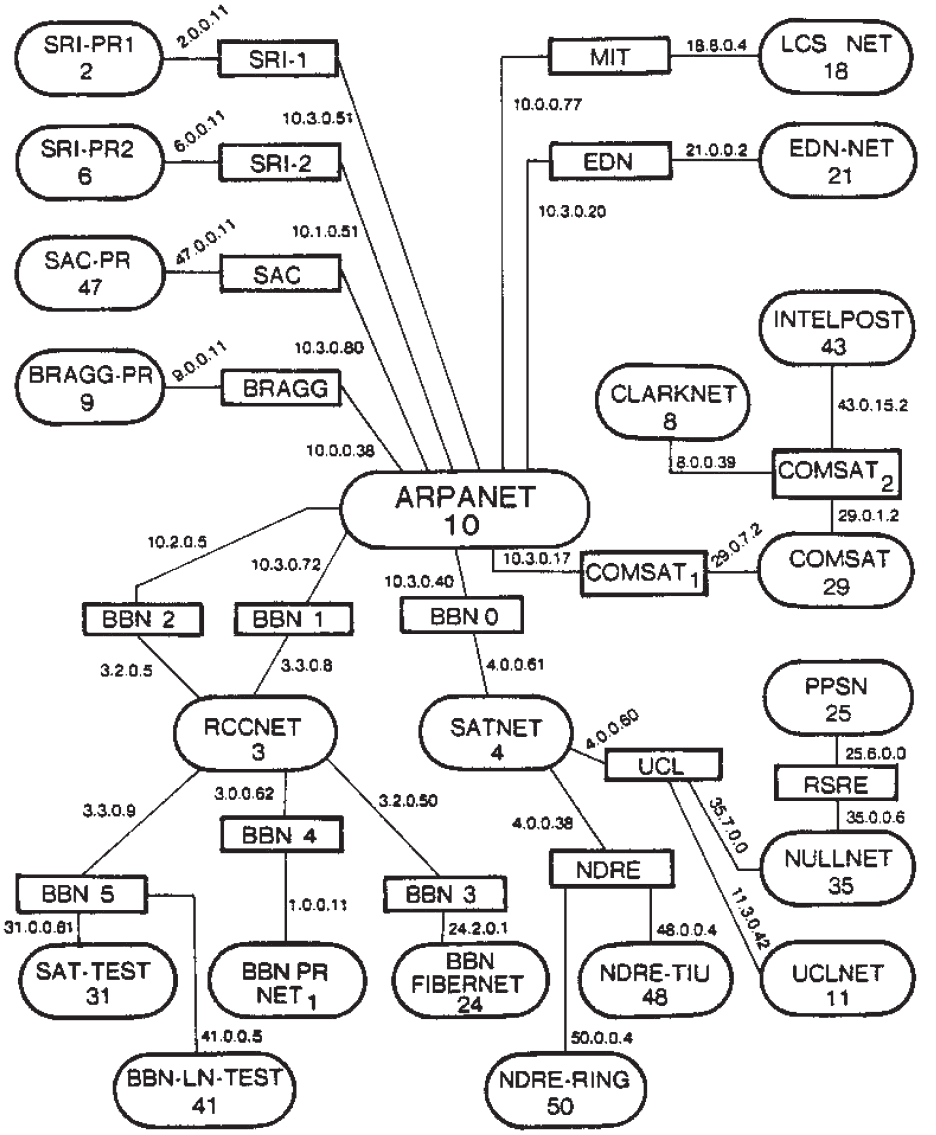
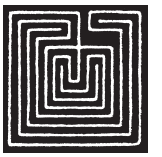
Dædalus

Journal of the American Academy of Arts & Sciences

Winter 2016

The Internet


SRI-PR1
2
2.0.0.11
SRI-1
SRI-PR2
6
6.0.0.11
SRI-2
SAC-PR
47
47.0.0.11
SAC
BRAGG-PR
9
9.0.0.11
BRAGG
10.3.0.51
10.1.0.51
10.3.0.80
10.0.0.38
ARPANET
10
MIT
18.6.0.4
LCS NET
18
10.0.0.77
EDN
21.0.0.2
EDN-NET
21
10.3.0.20
INTELPOST
43
43.0.15.2
CLARKNET
8
8.0.0.39
COMSAT 2
29.0.1.2
COMSAT
29
29.0.7.2
COMSAT 1
10.3.0.17
10.2.0.5
10.3.0.72
10.3.0.40
BBN 2
BBN 1
BBN 0
3.2.0.5
3.3.0.8
4.0.0.61
RCCNET
3
SATNET
4
4.0.0.60
UCL
PPSN
25
25.6.0.0
RSRE
35.0.0.6
NULLNET
35
35.7.0.0
3.3.0.9
3.0.0.62
3.2.0.50
BBN 4
4.0.0.38
NDRE
BBN 5
BBN 3
1.0.0.11
24.2.0.1
48.0.0.4
31.0.0.61
SAT-TEST
31
BBN PR
NET 1
BBN
FIBERNET
24
NDRE-TIU
48
11.3.0.42
UCLNET
11
41.0.0.5
50.0.0.4
BBN-LN-TEST
41
NDRE-RING
50



Inside front cover: Map of the entire Internet as of February 1982, showing the major component networks that made up the Internet at the time: the ARPANET (Advanced Research Projects Agency Network) at the center of the picture; the satellite network (SATNET), which hooked the United States to England and Norway; the early packet radio network PRNET; as well as the early host sites. Some of the original network numbers are still in use. The map was drawn by Jon Postel of the Information Sciences Institute, under a DARPA (Defense Advanced Research Projects Agency) research contract supporting Internet development.

Dædalus

Journal of the American Academy of Arts & Sciences

The labyrinth designed by Dædalus for King Minos of Crete, on a silver tetradrachma from Cnossos, Crete, c. 350–300 B.C. (35 mm, Cabinet des Médailles, Bibliothèque Nationale, Paris). "Such was the work, so intricate the place, / That scarce the workman all its turns cou'd trace; / And Dædalus was puzzled how to find / The secret ways of what himself design'd." – Ovid, *Metamorphoses*, Book 8

Dædalus was founded in 1955 and established as a quarterly in 1958. The journal's namesake was renowned in ancient Greece as an inventor, scientist, and unriddler of riddles. Its emblem, a maze seen from above, symbolizes the aspiration of its founders to "lift each of us above his cell in the labyrinth of learning in order that he may see the entire structure as if from above, where each separate part loses its comfortable separateness."

The American Academy of Arts & Sciences, like its journal, brings together distinguished individuals from every field of human endeavor. It was chartered in 1780 as a forum "to cultivate every art and science which may tend to advance the interest, honour, dignity, and happiness of a free, independent, and virtuous people." Now in its third century, the Academy, with its more than five thousand members, continues to provide intellectual leadership to meet the critical challenges facing our world.

Dædalus Winter 2016
Issued as Volume 145, Number 1

Editorial offices: *Dædalus*, American Academy of Arts & Sciences, 136 Irving Street, Cambridge MA 02138. Phone: 617 576 5085. Fax: 617 576 5088. Email: daedalus@amacad.org.

Library of Congress Catalog No. 12-30299.

Dædalus publishes by invitation only and assumes no responsibility for unsolicited manuscripts. The views expressed are those of the author(s) of each article, and not necessarily of the American Academy of Arts & Sciences.

Dædalus (ISSN 0011-5266; E-ISSN 1548-6192) is published quarterly (winter, spring, summer, fall) by The MIT Press, One Rogers Street, Cambridge MA 02142-1209, for the American Academy of Arts & Sciences. An electronic full-text version of *Dædalus* is available from The MIT Press. Subscription and address changes should be addressed to MIT Press Journals Customer Service, One Rogers Street, Cambridge MA 02142-1209. Phone: 617 253 2889; U.S./Canada 800 207 8354. Fax: 617 577 1545. Email: journals-cs@mit.edu.

Printed in the United States by The Sheridan Press, 450 Fame Avenue, Hanover, PA 17331.

Newsstand distribution by Ingram Periodicals Inc., 18 Ingram Blvd., La Vergne TN 37086.

Postmaster: Send address changes to *Dædalus*, One Rogers Street, Cambridge MA 02142-1209. Periodicals postage paid at Boston MA and at additional mailing offices.

The typeface is Cycles, designed by Sumner Stone at the Stone Type Foundry of Guinda CA. Each size of Cycles has been separately designed in the tradition of metal types.

Subscription rates: Electronic only for nonmember individuals – $48; institutions – $133. Canadians add 5% GST. Print and electronic for nonmember individuals – $54; institutions – $148. Canadians add 5% GST. Outside the United States and Canada add $23 for postage and handling. Prices subject to change without notice. Institutional subscriptions are on a volume-year basis. All other subscriptions begin with the next available issue.

Single issues: $14 for individuals; $37 for institutions. Outside the United States and Canada add $6 per issue for postage and handling. Prices subject to change without notice.

Claims for missing issues will be honored free of charge if made within three months of the publication date of the issue. Claims may be submitted to journals-cs@mit.edu. Members of the American Academy please direct all questions and claims to daedalus@amacad.org.

Advertising and mailing-list inquiries may be addressed to Marketing Department, MIT Press Journals, One Rogers Street, Cambridge MA 02142-1209. Phone: 617 253 2866. Fax: 617 253 1709. Email: journals-info@mit.edu.

Introduction

Yochai Benkler & David D. Clark

The Internet was born in 1983. At least that is when it adopted TCP/IP (transmission control protocol/Internet protocol), the communications protocol that conceptually separates the Internet from its predecessors and continues to define it, technically, to this day. It was originally designed in an academic environment, funded by the kinds of deep public research funds on which basic science depends. The direction and shape of research was left largely in the hands of researchers, and they built a system only a researcher could love: general, abstract, optimized for nothing, and open to exploration of more or less anything imaginable using connected computers. Thirty-two years later, the Internet has become the most fundamental global communications and knowledge infrastructure of our age, and is fast becoming the basic data-and-control network of the coming decade. It has evolved over its thirty-two years from a network that primarily delivered email among academics and government employees, to a network over which the World Wide Web arose, to the video and mobile platform it has become – and the control network for embedded computing that it is fast becoming.

Could the Internet have been different? Could it still evolve into a fundamentally different platform than what we have grown accustomed to? What design choices did designers make that resulted in the Internet as we know it, and what design choices are we currently making that will shape it in the future? The essays we compiled for this volume represent an effort to offer some insight to both the research community and society at-large about what is at stake in this discussion and what different choices imply about the fu-

doi:10.1162/DAED_x_00360

ture of the Internet. The ambition of this collection is not to cover the entire gamut of challenges and design choices, but to offer a richly detailed exploration of the kind of analysis required across the board.

Our own individual essays in the issue frame this discussion, presenting broad definitions of the challenges and exploring the problems and opportunities they entail. In "The Contingent Internet," David Clark reviews the history of the design choices that made the Internet what it became, and outlines the range of design choices we are likely to face in the coming years. Yochai Benkler outlines, in "Degrees of Freedom, Dimensions of Power," the ways in which the first-generation Internet diffused economic, social, and political power, and the series of changes that has created new control points around which both nation-states and market actors are concentrating power and creating new design challenges.

Five subsequent essays dive deeper into particular design challenges presented by the emerging Internet. Peter Kirstein elaborates on what it would take, technically, to build an Internet capable of scaling to the billions, or perhaps trillions, of nodes that the "Internet of Things" (in short, sensors everywhere) will require. Deborah Estrin and Ari Juels examine, in "Reassembling Our Digital Selves," what design elements could make the power of ubiquitously collected data safely available to individuals as "small data," rather than emerging purely as "big data" analysis for the use of larger entities.

In "Choices: Privacy and Surveillance in a Once and Future Internet," Susan Landau examines the deep concerns about security and privacy, for both society and individual, that have pushed to the fore in our increasingly connected world. She then outlines the design choices that would make it possible to attain both values in the teeth of trends that seem to offer neither. In "As Pirates Become CEOs," Zeynep Tufekci examines the displacement of the public Internet by increasing reliance on proprietary networks, like Facebook, for the most basic communications capabilities. She identifies the new opportunities for manipulating consumer demand and political action, and discusses the power shift that lapses in Internet security cause and the stresses that an advertiser-supported Internet places on the open Internet. Finally, John Palfrey explores the "Design Choices for Libraries in the Digital-Plus Era," and the stakes of these design choices for the role of publicly spirited organizations in an increasingly privately owned networked environment, individualized and abstracted from place.

Several core themes emerge from the efforts of these seven essays to define the design challenges we face in the coming years of Internet evolution:

The technical is political. As Clark's framing makes clear, even in the early days of the Internet, designers understood that design choices had political influence, particularly at the level of recognizing potential tensions between large computer providers such as IBM and the telecommunications carriers. Three decades later – and after two decades of consistent work in law, philosophy, and social science – the secret is out: the technical is the political. Different design choices are subject to conflict among governments, corporate stakeholders, and Internet users, all of whom pursue power and their (at times conflicting) interests through these choices.

Both Clark and Benkler's essays provide a rich description of how design choices affect ethical and political values in concrete settings. In their contribution, Estrin and Juels very clearly explore the tensions between design choices that are conducive to "big data" – the collection of information by large data processors seeking to learn about, and thereby influence, their users or customers – and those design choices that would be conducive to "small data." The latter decisions could empower users to ac-

cess their personal data and use it to manage their own lives, as well as gain personal services that would not be possible otherwise, but at the risk of personal data exposure and against the challenge of wresting control of small data away from companies pursuing big data capabilities. Landau, in turn, outlines the tensions between creating an Internet system that prioritizes individual privacy and safety, and a digital environment that may be secure, but that nonetheless makes users vulnerable to the surveillance systems of service providers and governments. Tufekci explores how the design characteristics of different social networks influence the type of communication feasible: she describes how the different designs of Twitter and Facebook caused the two platforms to diverge in the degree to which their algorithms directed attention to recent political protests in Ferguson, Missouri; and how the shift to Facebook from open Internet blogs changed the nature of online publication in the Iranian and Egyptian public sphere. She then further examines how algorithms can influence users' political and economic preferences, with substantial implications on both economy and democracy.

Smartphones and Things. A second major theme that emerges from this collection of essays is that the nature of the endpoints of the network has changed radically since the early days of the Internet, and this, in turn, has changed the design choices and their implications. The early Internet connected general purpose computers that were fixed in location and often shared among users. A node was not a person, but a computer, and a computer was a general purpose device, not a specific appliance connected for control. Today, the majority of Internet users connect using smartphones, which are both personal and mobile. Kirstein's essay wrestles with the substantial challenges of building a network intended to serve over one trillion devices, many of them special purpose machines aimed at sensing and control systems, without substantial embedded intelligence. The essays by Estrin and Juels and by Landau both attempt to address problems that come from the fact that ubiquitous connected computing also functions as a pervasive surveillance and control network.

The privatized Internet creates new challenges, in particular for the continued role of public institutions. Palfrey's essay presents libraries as a microcosm of a much broader problem that the Internet has created. Like Tufekci, Palfrey starts with the fact that the platforms that most people use to access the Internet are privately owned. What role, then, do public institutions have in this privatized environment? Are they obsolete? Are they a necessary counterweight to an increasingly privatized space? Palfrey eloquently argues for the continued vitality and essential role of public institutions in a thriving society. Meanwhile, Landau relies extensively on private actors, both market and nonmarket, to build privacy and security measures to protect users from both other market actors and nation-states. Estrin and Juels, by contrast, emphasize the consistent tensions that technical solutions alone create, and challenge us to design mixed legal, technical, and ethical frameworks to achieve a privacy characterized by what media analyst and computer scientist Helen Nissenbaum has called "contextual integrity." Understanding the role of public institutions and values in shaping the privately owned open spaces of the Internet will continue to be a major challenge in the coming years.

Actionable data. Several of the essays raise the prospect of increasingly actionable data becoming the core utility of the Internet. For the Internet of Things, it is easy to see. Data is no longer merely for monitoring, it is also applied for automatic control of the behavior of connected devices. Who owns these data and how they are secured so that unauthorized actors do not have the capac-

ity to act maliciously from a distance are central to the questions of security, privacy, and control throughout the issue. No less important, a mixture of data-analysis techniques and the personalized data available from Internet use today makes data about individuals actionable. Estrin and Juels seek to make small data actionable for individuals, and for their benefit. Tufekci examines how platforms combine data, behavioral sciences, and platform algorithms to predict and manipulate users' actions, perceptions, and emotions. Landau explores how users can protect themselves from being monitored and even acted against based on acquired data.

The influence of advertising. As both Clark and Tufekci emphasize, several of the core utilities of the Internet – Google and Facebook most prominently – depend on advertising to fund their operations. As a result, these advertising-supported services are developing the model of widespread surveillance and the use of actionable data to shape the Internet experience of their users, thereby increasing targeted purchasing. Many of the core tensions around privacy and between public and private values are a function of the fact that consumers demand "free" services, which providers develop and support only through the sale of user information to advertisers. As long as these core Internet utilities are privately provided and depend on advertising, the pressures on privacy and the tensions between providers and users will remain. Unless we find a way to allow users to pay for these utilities, this tension will remain at the core of design choices about how services are delivered, how much autonomy users have, and how much providers will be able to control and monetize the behavior of users.

The Internet started its life as public infrastructure, largely dedicated to communications among academic and public institutions. Over time, it turned into the core communications and information infrastructure of a networked economy and society. And it is now rapidly developing as a control system and organizational platform for the physical environment, through the Internet of Things, and is becoming ever more tightly integrated with the daily flow of life for individuals through mobile and wearable computing. In these transitions, it has become increasingly privately owned, commercial, productive, creative, and dangerous. It has become indispensable to an ever growing range of human activity. Understanding the design challenges these changes pose, subjecting them to continuous critical reflection informed by real-world analysis of the rapidly changing character of the Internet, and insisting on open, rational, democratic debate over the implications of our choices is perhaps the most important role of academic reflection about the past and future Internet.

YOCHAI BENKLER is the Berkman Professor of Entrepreneurial Legal Studies at Harvard Law School, and serves as Faculty Co-Director of the Berkman Center for Internet and Society at Harvard University. He is the author of *The Wealth of Networks: How Social Production Transforms Markets and Freedom* (2006), which won awards from the American Sociological Association and the American Political Science Association.

DAVID D. CLARK, a Fellow of the American Academy since 2002, is Senior Research Scientist at the MIT Computer Science and Artificial Intelligence Laboratory. He has been involved in the design of the Internet since the mid-1970s, and is a member of the Internet Hall of Fame. His recent policy publications include a chapter in *Trust, Computing, and Society* (ed. Richard H. R. Harper, 2014), and articles in the journals *Telecommunications Policy* and *Journal of Information Policy.*

The Contingent Internet

David D. Clark

Abstract: The Internet is so omnipresent and pervasive that its form may seem an inevitability. It is hard to imagine a "different" Internet, but the character of the Internet as we experience it today is, in fact, contingent on key decisions made in the past by its designers, those who have invested in it, and those who have regulated it. With different choices, we might have a very different Internet today. This paper uses past choices made during the emergence of the early Internet as a lens to look toward its future, which is equally contingent on decisions being made today: by industry, by governments, by users, and by the research community. This paper identifies some of those key choices, and discusses alternative futures for the Internet, including how open, how diverse, how funded, and how protective of the rights of its users it may be.

DAVID D. CLARK, a Fellow of the American Academy since 2002, is Senior Research Scientist at the MIT Computer Science and Artificial Intelligence Laboratory. He has been involved in the design of the Internet since the mid-1970s and is a member of the Internet Hall of Fame. His recent policy publications include a chapter in *Trust, Computing, and Society* (ed. Richard H. R. Harper, 2014), and articles in the journals *Telecommunications Policy* and *Journal of Information Policy*.

Is it possible that the Internet might never have happened? Is it possible that, in a parallel universe where the Internet's inventors had pursued different careers, we could be without a network that links all of our computers together? That we might have "personal computers" that were truly personal, not connected to the larger world unless their contents were copied to disk and mailed?

Actually, that alternative outcome is highly improbable. The Internet was in some respects a creation of its time: in the 1960s, the idea of a global network for computers was "in the air." A visionary of the time, J. C. R. Licklider, had already predicted teleconferencing, information sharing, instant messaging, online tax preparation, offshoring, and the potential for a digital divide.[1] However, at the time of the Internet's launch, there were competing conceptions for how to build a "computer network." Our alternate universe is not without the Internet, but rather is with a very *different* Internet.

This possibility may itself seem surprising: the Internet today is so omnipresent, so much a fixture of our lives that it seems almost as if it "had to be that way." What might an alternate Internet have looked like? This is an important question, because to recognize that there were multiple options for the early Internet, and that the Internet as we know it is contingent

doi:10.1162/DAED_a_00361

on decisions that could have led to different outcomes, is to recognize that the future of the Internet is itself contingent. Society will meet forks in the road that will determine the future of the Internet, and recognizing these points and discussing the alternatives, rather than later looking back and wondering if we chose the right path, is an opportunity we cannot forego.

The Internet is a "general purpose" network, designed for a variety of uses. It is suited to email, watching video, playing a computer game, looking at Web pages, and myriad other applications. To an Internet engineer, the Internet is the system that moves data, and the applications (like a Web browser, which users might lump into the larger concept of "Internet") run on top of that data-transport service. This modularity, and this generality, seem a natural way to structure a network that hooks computers together: computers are general-purpose devices; since the Internet hooks computers together, it too ought to be general. But this idea was quite alien to the communications engineers of the early-Internet era, who largely worked for telephone companies. They asked what was to them an obvious question: how can you design something if you don't know what it is for? The telephone system was designed for a known purpose: to carry telephone calls. The requirements implied by that purpose drove every design decision of the telephone system; thus, the engineers from the world of telephone systems were confounded by the task of designing a system without knowing what its requirements were. The early history of the Internet was therefore written by people who came from a computing background, not a classical network (telephone) background. Most computers are built without a singular purpose, and this mind-set drove the Internet's design.

But this generality has a price. The service the Internet delivers is almost certainly not optimal for any particular application. Design for optimal performance and design for generality are two distinct objectives. And it may take more effort to design each application in a general network than in a network that was tailored to each application. Over the decades of the Internet's evolution, there has been a succession of dominant applications. In the early years of the Internet, the Internet was equated to email, and to ask people if they were "on the Internet" was to ask if they had an email address. Email is an undemanding application to support, and if the Internet had drifted too far toward exclusively supporting it (as was happening to some degree), the Web might not have been able to emerge. But the Web succeeded, and its presence as a complement to email reminded engineers of the value of generality. But this cycle repeats, and the emergence of streaming audio and video in the early 2000s tested the generality of an Internet that had drifted toward a presumption that now the Web, and not email, was *the* application. Today, streaming, high-quality video drives the constant reengineering of the Internet, and it is tempting once again to assume that we know now what the Internet is best suited for, and optimize it accordingly. The past teaches us that we should always be alert to protect the generality of the Internet, and allow for the future even when faced with the needs of the present.

There is another aspect of generality: the applications that run over the basic transport service of the Internet are not designed or distributed by the same entity that provides the basic data-transport service. This characteristic has been called the "open" Internet, and again, this separation made sense to a computer engineer but did not fit conceptually with the telecommunication engineer. The telephone company installed that wire to your house to sell you telephone service, not to enable some other company to sell you theirs. From the telephone com-

pany's perspective, it is expensive to install all those wires, and how could they get a reasonable return on investment if they were not the exclusive service provider?

In the early days of the Internet, the only way to access the Internet from home was to use a modem to make a dial-up connection to an Internet service provider (ISP). A residential user paid the telephone company for the telephone service, and then paid the ISP for providing access. This seemed then like a minor shift in the business model of the telephone companies. But as the possibility of expanding broadband services to the home emerged in the 1990s, the corporate resistance to an open platform became quite clear. One telephone executive explained to me at the time: "If we don't come to your party, you don't have a party. And we don't like your party very much. The only way you will get broadband to the home is if the FCC forces us to provide it."

That was a fork in the road, and the Internet certainly might have taken another path. In fact, the force that led the Internet toward residential broadband was, to a considerable extent, the emergence of the cable television industry as a credible and competitive provider of high-speed residential Internet.

We continue to see echoes of this tension between the Internet as an open platform for third-party applications and broadband access as an expensive investment that should work to the advantage of its owner. The current debates around the concept of "network neutrality" are at their heart about whether broadband providers should be regulated to provide a neutral, open platform for third-party services, or if they have the right to define the services they offer (and perhaps favor) over the infrastructure they invested in building.

Another consequence of generality is that the data-transport layer of the Internet has no concept of what the application is trying to do (as opposed to the design of the telephone system, which at all levels reflects the centrality of the telephone call). If the design of the Internet required that the network understand what the application were doing, deploying a new application would require its designer to somehow modify the core of the network to include this knowledge. To the early designers, this was a fork in the road down which they did *not* want to go. If an application designer had to alter the network before deploying a new application, this would both complicate the process of innovation and create potential for the network to block one or another application.

The Internet has been called the *stupid network*, the telephone system being the *intelligent network*; the open-design approach of the Internet makes perfect sense – that is, until things go wrong. If the network itself is impairing the operation of an application, the network cannot always detect or correct this. The network may be able to detect that one of its components has failed, but more complex failures may go undetected, leaving frustrated users who can see that their application is not working, but who have no remedy available to them. Had we taken the fork in the road that enabled the network to know more about what each application was trying to do, the network might have been less supportive of easy innovation, but might also have been less frustrating to use when unexpected problems inevitably arose.

Finally, the division of responsibility between the provider of the data-transport service and the provider of the application means that responsibility for core requirements like security is divided among several actors. This both makes the objective harder to achieve and adds incentive to delegate the task to another party. In this way, the design decisions that shaped the Internet as we know it likely did not optimize secure and trustworthy operation.

These design choices led to differences in the technical character of the Internet, but many choices also led to particular outcomes in the industrial structure of the Internet ecosystem. When we made design decisions about system modularity in the early Internet, it was not entirely clear to us that our design was both a technical structure and an industrial structure. Some of the early network pioneers, though, certainly did understand this. In the 1970s, there was a substantial debate between advocates of two sorts of networks: *datagram* and *virtual circuit*. Datagram networks have a simpler core, with more functions shifted to hosts at the edge. Virtual-circuit networks have more function in the core of the net, and thus more power and control shifted to the network operator. The Internet is a datagram network; the ARPANET (Advanced Research Projects Agency Network) that preceded it was a virtual-circuit network.

One of the most vocal advocates of the datagram approach was the French computer systems designer Louis Pouzin, who was building a datagram network called Cyclades at the same time that the Internet was taking shape. In 1976, he published a paper that reached the following conclusion:

> The controversy DG vs. VC in public packet networks should be placed in its proper context.
>
> First, it is a technical issue, where each side has arguments. It is hard to tell objectively what a balanced opinion should be, since there is no unbiased expert. This paper argues in favor of DGs, but the author does not pretend being unbiased. Even if no compromise could be found, the implications would be limited to some additional cost in hardware and software at the network interface. So much resources are already wasted in computing and communications that the end result may not be affected dramatically.
>
> Second, the political significance of the controversy is much more fundamental, as it signals ambushes in a power struggle between carriers and computer industry. Everyone knows that in the end, it means IBM vs. Telecommunications, through mercenaries. It may be tempting for some governments to let their carrier monopolize the data processing market, as a way to control IBM. What may happen, is that they fail in checking IBM but succeed in destroying smaller industries. Another possible outcome is underdevelopment, as for the telephone. It looks as if we may need some sort of peacemaker to draw up boundary lines before we all get in trouble.[2]

Pouzin saw the battle over control of the global network as a battle between the computer industry and the telecommunications industry. At the time, the computer industry was dominated by huge players like IBM, giving shape to Pouzin's "Battle of the Titans." IBM was a vertically integrated corporation, just as the telephone companies were: if a firm got its hardware from IBM, it likely got its software from IBM as well. Pouzin may not have foreseen the coming shift in the computer industry to support more open-hardware platforms, but he clearly saw different technical decisions as shifting the balance of power from one industry sector to another.

In contrast to the Internet, Pouzin's Cyclades network was ultimately unsuccessful. Its failure is (speculatively) attributed to the hostility and resistance of the French PTT (postal, telegraph, and telephone government unit).

One of the lessons of the past is that the users of the Internet are an active force in defining what the network is, both by choice of application and by the creation of unexpected applications never anticipated by network engineers. This trend continues today with the success of user-created programs for peer-to-peer music sharing, for example. Sometimes users take the network down a fork in the road that the designers

did not contemplate, or perhaps had even dismissed. In the very early days of computer networking, the designers were focusing on remote access to the expensive, high-power computers of the time. One of the early network engineers asserted that message services were "not an important motivation for a network of scientific computers."[3] Of course, users proved him wrong, flocking to email en masse.

The second, related lesson is that the open character of the Internet is what allows this sort of user-driven evolution to take place. In a more structured and vertically integrated vision of computer networking, the network provider might not even choose to offer an email application. The Internet, by its structure, is amazingly open to exploration by users and third-party innovators. This benefit, though, was clearly contingent on earlier design choices.

Is the core technology and structure of the Internet set for the indefinite future? Or are there further forks in the road that might change the basic character of the Internet? One important and ongoing debate concerns the extent to which encryption should be used by default to protect communication between users from observation (and modification). If the goal is privacy of communication among communities of users, encryption is a powerful tool. But encryption thwarts the goals of many other actors: the intelligence community benefits greatly from being able to spy on content, and in some nations, this right of spying is not subject to debate. If the Internet were to move to a posture of "encryption by default," would certain nations opt out of the public Internet as we know it and essentially create a separate, state-controlled network for their citizens? Further, Internet service providers find encryption between communicants problematic, since it prevents them also from seeing what their users are doing. If the Internet were still the totally open, neutral platform that only moved data from sender to receiver, such peeking would seem unnecessary; but that interpretation of today's Internet is an oversimplification. Operators claim that they need to see what users are doing in order to optimize their experience – which may at times be true – and to selectively influence what they are doing. Many of these interventions by Internet service providers – such as the modification of data in transit to insert advertisements – have been met with protest, but other interventions – like reformatting data to fit content onto the small screen of a mobile device – are more easily justified by users. The tussle over the use of encryption is only the latest chapter in the struggle between the providers of the data-transport service of the Internet, the providers of higher-level services and applications, the users, and the state system for control of the user experience.

The experience of using the Internet is becoming more diverse. While the data-transport service is more or less uniform across the globe, that service does not define the user experience. The experience of the user is defined by the applications that run "on top of" that service. If different applications are available or are preferred by users in a given region, the resulting Internet experience will, in turn, be different. Perhaps the most obvious example of this today is the Internet experience in China, where the state government has blocked access to many of the applications that define the Western user experience, such as Facebook and Twitter. But there are domestic equivalents to these services within China that make the Chinese Internet experience a vibrant space of interaction, even if it is heavily policed. Still, there is no easy way for a Chinese and American to "friend" each other using Facebook.

The global community will need to decide the extent to which we fight against this diversification of the user experience. Some of the early Internet visionaries con-

ceived the Internet as a platform for global discourse and a vector for a global civil society. Diversification of the Internet experience would seem to erode that vision. On the other hand, differences of language, culture, and norms are real, and would suggest that as the Internet matures, the experience in any region would evolve to conform to those factors. The generality of the platform does not mean that everyone has to use it the same way. In fact, the generality, combined with the ability of users to vote (via their usage) for the applications of their choice, almost makes regional diversity inevitable. If the nations of the world were to push for some sort of global alignment of regulation and incentive, it would likely lead to a more homogenous but less satisfactory Internet experience. Perhaps the ideal is an Internet that accepts the diversity of experience for most users, but permits interaction on a global level among users who seek it. We should urge China not to block Facebook (which they view as a threat to regime stability) but we should at the same time accept the outcome that most Chinese users prefer their domestic alternatives.

Another critical issue that will shape the future Internet is the poor state of Internet security. We hear almost daily about theft of data, computers corrupted with malicious software (malware), cyber-crime, and many other breaches of the security of both users and service providers. This state of affairs could play out in a number of ways. One is that the current state of insecurity persists, which might eventually prevent sensitive or important transactions from taking place online. In this way, poor security could be a barrier both to the uptake and utility of the Internet. An alternate future is that in the attempt to improve Internet security, the Internet mutates to hold users more accountable for their actions. For a number of reasons, the original design of the Internet did not include any mechanisms to deal with identity management. It was understood, if imperfectly, that different applications might call for differing degrees of accountability; for example, while a transaction between a customer and his or her bank calls for strong mutual verification of identity, another user may not be comfortable searching for information on AIDS, for example, if the query is not anonymous. Defining which actors can issue globally trustworthy identity credentials is, of course, another challenge to uniform identity management. These concerns remain valid; but at the same time, the pressure to improve our ability to hold users accountable and deter malicious behavior may push toward stronger identity tracking. In many nations today, one must provide a national identity number to use the Internet; respect for anonymous action may only be a local preference, which could erode under global pressure for accountability.

The explosion of mobile devices signals another inflection point for the future of the Internet. The traditional industry narrative pitted application designers against the Internet service providers: application designers wanted an open, neutral platform on which to innovate and the Internet service providers wanted control over the services offered in order to effectively monetize the user experience. Computer manufacturers were seen as neutral in this dynamic, lacking a business model that bundled proprietary applications on the device. But the trajectory of the mobile device is very different: the makers of smart phones show a much greater interest in shaping (and monetizing) the user experience. Apple charges a fee (currently 30 percent) for the sale of a paid application for the iPhone or iPad through their app store. The regulators in the United States would be quick to intervene if ISPs tried to charge a fee to customers for using a specific app, but so far there has been little criticism of device makers doing exactly that.

The interplay of device maker, ISP, and application creator and provider is partic-

ularly interesting in developing nations. Getting developing states online is socially desirable; the power of connectivity to improve the conditions of citizens is evident. But what strategies are acceptable in pursuit of this goal? Here, Facebook has launched a clever scheme: they developed a stripped-down version of the Facebook application (called 0.facebook.com; the zero implies zero cost) and have negotiated an arrangement with mobile service providers in many developing countries ensuring that use of the application is free to users, not counting against any data quotas. By making use free – and in some cases even arranging for a discount on the device – the uptake of Internet in the developing world may increase. But as a consequence, a generation of users will equate the Internet not with sending email, not with searching the Web, but with using Facebook. This is not a hypothetical outcome; surveys suggest it is already happening.[4] Is this degree of corporate capture acceptable? And what growth potential does it limit? Of course, this is an issue that each country will decide for itself through domestic regulation (or deregulation, for that matter). But again, such decisions will likely diverge the character of the Internet experience as it evolves across the globe.

The design alternatives I have described may seem to concern principally the user experience, but at a deeper level, they are struggles over control. The previous examples illustrate tussles between ISPs, application designers, device makers, and governments (among others) over control of the Internet. Depending on how the balance of power among these actors evolves, we may see different outcomes with respect to deployment, openness, innovation, and user experience. The immense power of private-sector actors is notable; society has largely left the future of the Internet in the hands of far-reaching, profit-seeking entities. In the United States, there is a tendency to put our hope in competition, as if with enough competition the market will converge on what users prefer. Sadly, however much we wish, there is simply not much competition to build residential broadband access networks; the investment and risks required to become competitive are too great. And at the layer of applications, there seems to be a recurring tension between building an application that attracts users and developing an application that makes money, perhaps by capturing information about the user to be used in more selective (and thus more pricey) advertisements. Competition cannot discipline this behavior if the behavior alone allows commercial providers of applications to make money.

Thus, another fork in the road is how the Internet will be paid for. Today, aside from some public-sector money that supports specific challenges like rural deployment, there are only two important sources of money to pay for the Internet: the fees we pay as service subscribers, and advertising. Internet users today pay for broadband access, and they pay for an assortment of applications and services, including streaming video, online games, and music services. But advertising pays for the "free" Internet experience, those websites that cost users nothing to visit. And advertising dollars will not grow without bound. Spending on advertising can only be a fraction of commerce, or ecommerce, as is most common on the Internet. In 2013, $42.8 billion was spent on Internet advertising in the United States.[5] The total monthly advertising expenditure per household with broadband access (eighty-eight million households in 2013) is about $40.[6] In other words, all of the advertising-supported Internet content is fighting over an amount that is smaller than the average monthly cost of broadband access. Could the Internet experience stall because we run out of advertising dollars?[7]

Online advertising will indeed grow as it cannibalizes traditional TV advertising. But

what if a new "Internet experience economy" arises, in which users pay a small amount for access to a broad spectrum of Internet applications that do not track their usage and do not run advertisements? Users are largely accustomed to paying for both premium apps on mobile devices and music and video. Perhaps there will be a shift in how they pay for access to Web content. And were this to happen, who would control that payment ecosystem?

The next major tussle over control involves the governments of the world. The private sector has a common set of motivations: be profitable, grow, survive. Governments have a range of concerns: national security (which can include regime stability), law enforcement, taxation, control of "unacceptable" content, and protection of the rights of powerful private-sector actors (such as protection of copyright), among many others. Different countries have different priorities, different laws, and different approaches to governance. In some cases, these priorities put them at odds with the private sector that, in most countries, has a dominant influence over the character of the Internet. In late 2012, at the World Conference on International Telecommunications in Dubai, the International Telecommunications Union (a division of the United Nations) proposed an international treaty that would give it the right to regulate international interconnection in the Internet. This idea was supported by a number of powerful nations, but it failed to gain traction. However, this preference for state control over important aspects of the Internet will probably continue to grow in certain quarters.

But in all of this contention over the future of the Internet, there is one set of actors that has faded from view: the federally funded research community that designed and built the Internet. From one point of view, this trajectory is proper: they did their job, the commercial world has taken over, and the Internet is now an engine of economic innovation. But from another point of view, there might be a richer, more diverse set of uses for the Internet if nonprofit actors were motivated and supported to develop "noncommercial" applications. As the proliferation of apps for mobile devices suggests, it is not hard to launch a new application today. Perhaps one way to pick among future alternatives for the Internet is for interested parties to vote with their dollars, funding the development of applications that are not motivated by the pursuit of profits, but by interest in civic, cultural, or political participation.

It is possible that if we leave the future of the Internet in the hands of powerful private-sector players, we will get the outcome we want. It is possible that if we allow the governments of the world to make decisions that shape the future of the Internet, we will get the future we want. But there is great risk in being passive about the Internet's future; it may simply be too important to leave either to the forces of commerce or the mechanisms of global politics. Perhaps the most important question is how the voice of the users of the world can be injected into the decisions that will shape the future of the Internet.

ENDNOTES

1 J. C. R. Licklider and Robert Taylor, "The Computer as a Communication Device," *Science and Technology* (April 1968).

2 Louis Pouzin, "Virtual Circuits vs. Datagrams: Technical and Political Problems," *Proceedings of the June 7 – 10, 1976, National Computer Conference and Exposition* (New York: Association for Computing Machinery, 1976), 483 – 494.

3 Lawrence G. Roberts, "Multiple Computer Networks and Intercomputer Communication," *SOSP '67 Proceedings of the First ACM Symposium on Operating Systems Principles* (New York: Association for Computing Machinery, 1967).

4 Leo Mirani, "Millions of Facebook Users Have No Idea They're Using the Internet," *Quartz*, February 9, 2015, http://qz.com/333313/milliions-of-facebook-users-have-no-idea-theyre-using-the-internet/.

5 Interactive Advertising Bureau and PricewaterhouseCoopers, *IAB Internet Advertising Revenue Report: 2013 Full Year Results* (New York: Interactive Advertising Bureau and PricewaterhouseCoopers, 2014), http://www.iab.net/media/file/IAB_Internet_Advertising_Revenue_Report_FY_2013.pdf.

6 National Telecommunications and Information Administration, United States Department of Commerce, "Household Broadband Adoption Climbs to 74.2 Percent," June 6, 2013, http://www.ntia.doc.gov/blog/2013/household-broadband-adoption-climbs-724-percent.

7 Moreover, how will the growth of programs like Adblock, which allows its two hundred million users to surf the Web without seeing most conventional ads, affect advertising revenue?

Degrees of Freedom, Dimensions of Power

Yochai Benkler

Abstract: The original Internet design combined technical, organizational, and cultural characteristics that decentralized power along diverse dimensions. Decentralized institutional, technical, and market power maximized freedom to operate and innovate at the expense of control. Market developments have introduced new points of control. Mobile and cloud computing, the Internet of Things, fiber transition, big data, surveillance, and behavioral marketing introduce new control points and dimensions of power into the Internet as a social-cultural-economic platform. Unlike in the Internet's first generation, companies and governments are well aware of the significance of design choices, and are jostling to acquire power over, and appropriate value from, networked activity. If we are to preserve the democratic and creative promise of the Internet, we must continuously diagnose control points as they emerge and devise mechanisms of recreating diversity of constraint and degrees of freedom in the network to work around these forms of reconcentrated power.

YOCHAI BENKLER is the Berkman Professor of Entrepreneurial Legal Studies at Harvard Law School, and serves as Faculty Co-Director of the Berkman Center for Internet and Society at Harvard University. He is the author of *The Wealth of Networks: How Social Production Transforms Markets and Freedom* (2006), which won awards from the American Sociological Association and the American Political Science Association.

In March 2000, AOL tried to pull a program that two of its employees had released online twenty-four hours earlier. Gnutella was a peer-to-peer file sharing program, and AOL was concerned about copyright liability. But Gnutella was free software, and it had been released, along with its source code, under the GNU General Public License. Gnutella was quickly adopted and developed by diverse groups, becoming the basis for a range of peer-to-peer (P2P) networks that either used or improved upon its source code. Technical architecture, cultural practice, social production, market structure, and timing had prevented AOL from halting the development of Gnutella.

Fourteen years later, in February 2014, Apple's app store rejected a game that mocked North Korean leader Kim Jong Un. Apple already had a history of blocking applications of which it disapproved: cartoons that mocked President Obama, an app for browsing State Department cables on WikiLeaks, or a game that criticized the company's treatment of its workers in iPhone manufacturing processes. Initially, Apple had also forced Skype to block usage on 3G mobile networks, rejected the Google Voice app, and disabled Google Maps on the iPhone. Here developments en-

doi:10.1162/DAED_a_00362

abled Apple to exert power over users and developers in a manner that was simply impossible a decade and a half earlier: smartphones running over proprietary cellular networks, an operating system integrated with hardware that controlled what software is preloaded and made available, and an "app store" model of software distribution.

In 1993, *The New Yorker* published a Peter Steiner cartoon with the caption, "On the Internet, nobody knows you're a dog." By 2014, Maidan protesters in Kiev could receive text messages that read, "Dear subscriber, you are registered as a participant in a mass disturbance."[1] Whether Internet design ultimately will support a high degree of freedom, as was offered by the first generation Internet, or will evolve toward a system that amplifies power in the hands of the state and a concentrated class of private actors, is the central design challenge of the coming decade.

In its first quarter-century, "the Internet" was not only a technical system, but also an innovative organizational system; an institutional system pervaded by commons; a competitive market with low barriers to entry; and, finally, a zeitgeist, cultural habit of mind, or ideology, perhaps best captured by the saying from computer scientist and early architect of the Internet, David Clark: "We reject: kings, presidents and voting. We believe in: rough consensus and running code."[2] It is the integrated effect of all these dimensions that should properly be understood as the Internet in its first twenty-five years, and it is changes in several of these elements that underwrite the transformation of the Internet into a more effective platform for the reconcentration of power.

The introduction of the iPhone in 2007 marked the shift to handheld computing and ushered in a shift to proprietary, controlled devices, software, and networks. Amazon's Elastic Compute Cloud (EC2) – introduced in 2006 – created another potential point of control. The coming of age of advertiser-supported platforms and the emergence, in 2008, of "big data" as both a working concept and catchphrase marked a new drive to collect data and deploy it. Big data may ultimately allow a small number of companies – those large enough to control, access, and analyze sufficient data – to predict, shape, and "nudge" the behaviors of hundreds of millions of people. Since the mid-2000s, home broadband has been replicating some of telecommunications' older monopoly characteristics, while ever-higher speeds are shifting usage further toward streaming video. Consumer demand for high-grade commercial video services, most prominently Netflix, has in turn increased the pressure to implement technical control measures in basic infrastructure, capped by the adoption of Digital Rights Management (DRM) as a core component of HTML5 in 2014. Together, these changes have destabilized the diverse open systems that had made up what we thought of as the Internet.

The design of the original Internet was biased in favor of decentralization of power and freedom to act. As a result, we benefited from an explosion of decentralized entrepreneurial activity and expressive individual work, as well as extensive participatory activity. But the design characteristics that underwrote these gains also supported cybercrime, spam, and malice.

By *power*, I mean the capacity of an entity to alter the behaviors, beliefs, outcomes, or configurations of some other entity. Power, in itself, is not good or bad; centralization and decentralization are not good or bad, in and of themselves. Centralized power may be in the hands of the state (legitimate or authoritarian) or big companies (responsive and efficient or extractive), and decentralized power may be distributed among individuals (participating citizens, expressive users, entrepreneurs, or criminals) or

loose collectives (engaged crowds or wild mobs). To imagine either that all centralized power is good and all decentralized power is criminal and mob-like, or that all decentralized power is participatory and expressive and all centralized power is extractive and authoritarian is wildly ahistorical.

Internet architecture shapes power, and unlike in the early days, everyone knows this now. Because power often involves the capacity to reshape terms of engagement, we are seeing extensive efforts to lock and extend existing power. If one were naive enough to imagine that all efforts at centralization were aimed merely at taming the "bad" decentralization, one might be sanguine about the fact that governments and companies are pushing toward greater centralization. Further, if one is paranoid enough to imagine that decentralization necessarily resolves to mob rule, then a similar sanguinity is called for. But in the absence of these assumptions, we are left with the task of maintaining an Internet both open enough and resistant enough to power to allow, at least, continued contestation of decisions to create points of control in the networked environment. If we allow that power can be good or bad, whether centralized or decentralized, and that existing dynamics are tending toward greater centralization and stabilization of power, then we are left with a singular task: to design a system that will disrupt forms of power – old and new – as they emerge, and that will provide a range of degrees of freedom, allowing individuals and groups to bob and weave among the sources and forms of power that the Internet is coming to instantiate.

That the original TCP/IP protocol outlines an open, loosely coupled system is, at this point, trivial. The basic end-to-end design principle it instantiates allows any application developer to use the networking protocol to send its payload, whatever that is, to its destination, wherever that may be, on a best-efforts basis. The generality of the protocol disabled crisp identification of the nature of parties to a communication, and offered no control points through which an entity could exclude or constrain another discrete entity attempting to use it. While the Internet protocol itself was a critical element, it was not, by itself, sufficient to diffuse power.

What typified the first quarter-century of the Internet was an integrated system of open systems. These included: the technical standards of the Internet and the World Wide Web; the decentralized, open organizational models of the Internet Engineering Task Force (IETF) and the World Wide Web Consortium (W3C); and the competitive market structure for connectivity (the low cost of copper wire, subject to common carriage rules, resulted in over five thousand Internet service providers, or ISPs) and devices (PCs became a commodity item). These systems were complemented by widespread use of open, standards-based devices (such as PCs running software developed and distributed by a diverse range of entities); the emergence of commons-based production, particularly free and open-source software (FOSS); and the culture of openness and resistance to authority shared by most early users and developers of components of the Internet ecosystem and its core applications. Together, these created a system designed to resist the application of power from any centralized authority, whether it pertained to free speech or to free innovation without permission, which was very much at the core of the Internet's architectural design principles.

Several developments suggest that we are shifting to an Internet that facilitates the accumulation of power by a relatively small set of influential state and nonstate actors. While the Internet protocol itself remains

open, as does the IETF, other control points counter the dynamics of the early Internet.

The first is the emergence of smartphones and the iOS app store. By the middle of 2014, Internet access by smartphone had surpassed Internet access from desktops or laptops.[3] Handheld and tablet users overwhelmingly used apps, rather than browser-based Internet access (Internet access via apps constituted 88 percent of handheld use and 82 percent of tablet use), and the growth rate of desktop use was 1 percent per year, while mobile app use grew more than 50 percent. Unless something dramatic changes these trends, the future of conscious Internet use is based in handheld devices running apps. Moreover, as connected sensors and controllers (origin of the "Internet of Things" as a concept) become pervasive, an increasingly larger portion of Internet use will not be conscious at all. The general-purpose device – owned and managed by its user and capable of running any software from any source – will continue to serve the portion of the population particularly interested in preserving its computational autonomy and in executing more challenging and complex tasks. But, as legal scholar Jonathan Zittrain warned in 2008, the majority of Internet-mediated practice will be undertaken with devices that are either narrowly customizable appliances or controlled on the app store model.[4]

The primary source of constraint on the Apple app store's center of power is competition from Android. In principle, Android OS (operating system) phones can use app stores other than Google's, and relatively simple alteration of the default settings allows users to sideload apps without the app store. In practice, while reliable numbers are scant, it appears that most Android apps are downloaded from Google Play or Amazon's app store. Habits of use and consumer convenience seem to largely negate the effects of the technical feasibility of sideloading. Limits, if any, on the power of the app store owners come from market competition between iOS and Android, and – perhaps, to the extent these constraints exist and are, further, given voice in the organizational cultures of these companies – from internal ethical or cultural constraints imposed by Google or Apple insiders on what counts as acceptable applications of power.

The increasing importance of mobile wireless cellular networks as core Internet infrastructure and these networks' management models are a second control point for us to consider. Wireless carriers have organizational habits rooted in a controlled and optimized network model. The carrier controls what devices are permitted, and knows, manages, and bills all users and usage. Congestion management and quality of service were early initial requirements for these companies, and the use of auctions to allocate spectrum to wireless carriers meant that they saw the physical infrastructure as privately owned and integrated with carriage services. The models of wireless telephony – technical, legal ownership, engineering culture, and business practice – were fundamentally built to enable control by the owner and service provider so as to optimize a known set of services to known paying consumers. These characteristics stood in contrast to the Internet model, through which carriers were legally excluded from control over the network; users and usage were unknown and assumed unknowable; resilient best-efforts, not quality of service, were the core commitment; flexibility to unknown, new uses and users trumped optimization for known uses and users; and any network and open-standards-compliant device could be connected to the network on an equal basis.

The most obvious example of power that follows directly from the historical model of wireless telephony was AT&T's requirement that Apple prevent Skype from using cellular (as opposed to WiFi) data on the

iPhone. Similarly, when carriers impose data caps, but then exclude favored services from counting against those data caps, they nudge users to adopt the preferred applications. In both cases, ownership of the spectrum and the service, the concept of optimization, and the integration of use with known paying users permit the company to exert control over what users can do and what companies unaffiliated with the service providers can offer. The controlled infrastructure, even where built to support control by commercial providers, also facilitates greater control by government agencies. The NSA's collection of bulk metadata from U.S. phone providers offers an obvious example of the more systemic shift in power that this new, more centralized architecture enables.

Packet discrimination and the end of legacy telephone copper-wire as physical infrastructure for broadband form a third control point. The first generation of Internet access by the public took place over dial-up connections. Becoming an ISP required little more than a modem bank connected to a phone line for users to dial; providers numbered in the thousands. The move to cable broadband and DSL over telephone lines increased the complexity of providing service and reduced the number of potential competitors. The deployment of the cable broadband DOCSIS 3.0 standard after 2006 meant that, in the long term, no more upgrades to the copper-wire telephone infrastructure would do. Only fiber-to-the-home could compete with cable for speed. The substantial civil engineering costs of fiber, in turn, reintroduced natural monopoly economics into home broadband markets, making competition a relatively weaker source of discipline for providers.[5]

The practical implication of the death of copper was that the home broadband provider became a significant point of control. At no point was this clearer than in the net neutrality debates. Most prominently, from late 2013 to early 2014, Netflix, Comcast, and Verizon FiOS clashed over whether the carriers were slowing Netflix's service in order to extract payment for adequate service. Independent studies confirmed that the slowdown occurred at the peering point – where Cogent and Level 3, carriers that Netflix uses to carry its traffic, connected to the Comcast and Verizon networks – and was likely caused by business disputes, not technical issues.[6] The parties blamed each other; but for our understanding, the vital development is that the gateway to the home broadband connection has become a central point of control, over which large corporations struggle (to the detriment of both end-users and competitors in the cloud who are not party to negotiations).

The re-emergence of natural monopoly economics in home broadband leaves us with a market or regulatory design choice, not a technical design choice. Barriers to entry into the wired home broadband market will continue to be high in the foreseeable future, hampering the efficacy of market solutions. Regulation in a number of forms seems most likely to diffuse power; this will likely require a combination of utility regulation – interconnection and interoperability on nondiscriminatory terms – and net neutrality rules requiring nondiscrimination among applications and content.

The emergence of cloud computing – enabled by increased speed of communications and widespread adoption of mobile computing – forms a third vital control point. Increasingly, individuals and businesses run their computation and storage remotely, on large computing and storage clusters owned and managed by third-party providers. This shift allows firms to economize on capital expenditures, enhance robustness and security, and scale computation, storage, and applications more flexibly than provisioning their own capacity would permit.

Despite the obvious benefits of cloud computing to individual users and firms, the technology also has the effect of centralizing power. The now-iconic example is Amazon's decision, in 2009, to delete copies of George Orwell's *1984* and *Animal Farm* from users' Kindles. The company claimed that the books were uploaded to the Kindle Store by a company that did not have the rights to them. Because Kindles are clients to a cloud service that stores and delivers the e-books, Amazon was in a position to delete these unapproved editions unilaterally. The platform, content, and software providers for cloud services all retain technical control over the data and operations of the customer in ways that were simply impossible when data and software were stored locally on the end-user's owned machine. The inherent power concern is not only about what the owner of the cloud provider can do, but also what third parties can do given the concentration of data and software in a single spot. One of the many revelations made by Edward Snowden was that the National Security Agency (NSA) project MUSCULAR had compromised both Google and Yahoo cloud storage facilities to enable the NSA to collect millions of records from e-mails, text, audio, and video from these companies.

What is important here is not that the NSA acted improperly; it is that cloud computing shifted the locus of power. When the data and software of hundreds of millions of people exist or run in a single place, whoever can compromise and gain control over it – legitimately or illegitimately – can exercise power over these hundreds of millions of people, at least to the extent that the data and applications extend power over their users and subjects.

The fourth control point is big data and its uses in behavioral control. In 2014, the *Proceedings of the National Academy of Sciences* reported on an experiment that manipulated the number of positive and negative emotional expressions on users' Facebook news feeds, which correlated with increased expressions by the subjects, of similarly positive and negative emotional content.[7] In sum, people's moods could be altered through manipulation of their news feeds. These findings complemented an earlier Facebook-based study that showed that users who received social messages notifying them that their friends had voted were more likely to vote than users who received no such message, or who received informational messages (as opposed to social).[8] The effect size was small in both cases, but statistically significant. The implication was quickly identified by scholars concerned with the power of Facebook and other companies that both control data and can integrate it, altering the user experience.[9]

Big data collection and processing, combined with ubiquitous sensing and connectivity, create extremely powerful insights on mass populations available to relatively few entities. These insights, together with new computational methods, make up what we think of as "big data." As Zeynep Tufekci has explained, when these methods combine with widespread experimentation (as in the Facebook experiments), behavioral science that analyzes individuals in a stimulus-response framework, and increasingly on-the-fly personalization of platforms, platform companies can nudge users to form beliefs and preferences, follow behaviors, and increase the probability of outcomes with ever-finer precision. These form the foundation of what management scholar Shoshana Zuboff has called "surveillance capitalism."[10] As consumers become more precisely and individually predictable in their behavioral response to experimentally derived stimuli, and platforms become ever-more programmable at an individual level to obtain desired behavioral responses, the idea of individual "preferences" that are exogenous and preexist market relations, and whose satisfaction drives mar-

kets and produces "welfare," becomes incoherent. While the endogeniety of preferences has been a central theme of critiques of markets, at least since economist Thorstein Veblen's *Theory of the Leisure Class*, behavioral manipulation has never been scientifically studied and integrated into service design on such a mass scale as has become possible, and increasingly standard, in big data/surveillance-informed behavioral marketing.

As part of the president's response to the political uproar caused by the Snowden disclosures, the President's Council of Advisors on Science and Technology (PCAST) issued a report on big data. The PCAST report was remarkable in that it repudiated two of the primary approaches we had previously used to preserve privacy: consent and anonymization. Since the emergence of "email privacy" as an issue in the early 1990s, regulatory efforts, particularly in the United States, focused on notice of collection and consent by the data subject. But as the PCAST report put it: "Notice and consent creates a nonlevel playing field in the implicit privacy negotiation between provider and user. The provider offers a complex, take-it-or-leave-it set of terms, while the user, in practice, can allocate only a few seconds to evaluating the offer. This is a kind of market failure."[11] As for anonymization, PCAST found that "[a]nonymization is increasingly easily defeated by the very techniques that are being developed for many legitimate applications of big data. In general, as the size and diversity of available data grow, the likelihood of being able to re-identify individuals (that is, reassociate their records with their names) grows substantially."[12] Both kinds of obsolescence mark a centralization of power, from individuals to the smaller set of entities capable of setting the terms of standard contracts or collecting, purchasing, and processing sufficient amounts of the ambient data surrounding individuals to defeat efforts at self-protection through anonymization.

PCAST's core recommendation was to accept the futility of regulating data collection and processing and implement more rigorous regulations on uses of collected data. Having diagnosed that both the technical systems involved in anonymization and the market systems involved in consent and contracting cannot alone carry the weight of preserving the *desiderata* we associate with privacy, PCAST shifted the onus of protection to the legal system. But this recommendation is undermined by the fact that the report in which it appears is itself the result of public exposure of a widely perceived failure of legal oversight. The Snowden revelations exposed that the complexity and opacity of the national security establishment rendered legal oversight and control highly imperfect. And this imperfection is not unique to government entities. The literature – ranging from rational-actor modeling through organizational sociology and cognitive bias – tells us that formalized rules imposed externally by a regulatory body are likely to function as imperfectly and incompletely as the technological or contractual subsystems that PCAST rejected. (This could be the case for a number of reasons, whether individual self-interest and agency problems; the force of habits, processes, and routines; or the dynamics of groupthink and bureaucratic culture.) All of these systems are radically incomplete and flawed, and it will be exceedingly difficult for any one of them to carry the burden of reversing a power flow instantiated in the basic architecture of the interaction.

The Netflix effect, and the increased identification of content as culture, form the final new control point I will discuss here. In January 2014, author and activist Cory Doctorow wrote a short post on his website, "We Are Huxleying Ourselves Into the Full Orwell." Doctorow was commenting on

the possibility that the W3C would adopt a standard for HTML5 that implements Digital Rights Management (DRM) in the basic browser standard.[13] The W3C was then being pushed to do this by browser manufacturers Microsoft, Apple, and Google, who were, in turn, being pushed by Netflix, which demanded DRM to assure its capacity to prevent users from creating unauthorized copies of its licensed content. By May 2014, not only had the W3C adopted the DRM standard, but the Mozilla Foundation, developer of the leading FOSS browser, had bowed to the perceived necessity of enabling users to view Netflix and released its own implementation of the DRM standard for HTML5. Together, these events reflect both the shift in cultural power and erosion of one of the core institutional and organizational mechanisms that made the Internet a force for decentralization of social, economic, and cultural power.

These events implicate several of the core design features of the early Internet and the policy battles to make it more readily susceptible to control. First, DRM technologies are a perfect example of an effort to impose power through technology. The essence of these technical measures is to allow one entity, originally a copyright owner, to determine who may make what uses of digital objects protected by DRM. The point is not legitimacy or legality, but power. DRM may be used equally to prevent unauthorized copying or to prevent legitimate fair uses of, or permissible innovation with, the encrypted materials. DRM technologies are designed to remove practical capacity to make a judgment about the legitimacy of a use from the possessor of the materials, and to locate that power with the copyright owner.

Although the U.S. Congress passed the Digital Millennium Copyright Act (DMCA) in 1998, which prohibited DRM circumvention, circumvention practices and devices have been trivially available to anyone who has chosen to use them. The practical capacity of copyright holders to control circumvention was nonexistent for music, and marginal for video. The adoption of DRM for video streaming as part of HTML5 sees the Web, one of the core open standards underlying a major use of the Internet, embed the control mechanism within it. The process of doing so exemplified an increasing role for major companies in the governance of standards, which had previously been more anarchic. Moreover, the adoption occurred due to widespread consumption patterns that put the Mozilla Foundation, a nonprofit organization dedicated to coordinate a FOSS project, in the position of either implementing a version of DRM or losing user share and becoming marginalized. It therefore suggests that the shift to widespread passive consumption usage patterns weakens the role that FOSS development could play to provision a separate, power-diffusing alternative infrastructure. The result is not only the singular decision to implement a particular technology; it is diagnostic of basic pressures created when the Internet intersects with mass media culture.

If commercial video is so important, what can we make of the claimed democratizing effect of Internet culture? Nielsen surveys suggest that watching video on the Internet represents about one-third of the amount of personal computer Internet use time for eighteen- to thirty-four-year-olds, about one-quarter for thirty-five- to forty-nine-year-olds, and about 15 percent for fifty- to sixty-four-year-olds.[14] Video on smartphones represented a smaller category of use. Imperfect measures, such as the relatively large share of Internet bandwidth consumed by Netflix in North America (about 35 percent),[15] and the high and growing rates of Netflix subscriptions among North American Internet users (rising from 31 percent to 38 percent of U.S. consumers from 2012 to 2013)[16] reflect the growing

significance of passive watching of professionally produced video entertainment online. Perhaps we are observing a shift toward using the Internet in ways more reminiscent of mass media than of the more culturally decentralized manner celebrated in the middle of the last decade, when fan videos and remixes were all the rage. Data from the Pew Research Center have suggested otherwise.[17] The proportion of adult American Internet users who have uploaded videos more than doubled from 2009 to 2013, reaching about one-third of Internet users. About 18 percent of users uploaded videos they produced for others to watch. Almost three-quarters of American adults online watch videos on YouTube, with comedy (57 percent), "how-to" (57 percent), educational (50 percent), and music videos (50 percent) being the most commonly viewed. These statistics suggest that while Internet users indeed seek Netflix and similar subscription services extensively, they also seek online video rooted in user-created, fan-shared videos. Importantly, the proportions of copyright-connected practices (comedy and music videos) and educational and free knowledge exchange ("how-to") videos are roughly similar.

From the perspective of cultural power, the rise of Netflix does not seem to imply displacement of distributed creativity. Rather, it occurs alongside continued expansion of decentralized cultural creation and decentralization of power, which can encourage, for instance, inserting memes and new frameworks into cultural discourse. Commercial platforms, like YouTube, Vimeo, and Flickr, developed to facilitate creation and distribution of culture by diverse users, offer one important pathway through controlled frameworks – like the app store on the handheld device – for continued sources of cultural decentralization to persist online. Nonetheless, the rise of proprietary video streaming as a major application seems to have been enough both to put pressure on the standards-setting process and to push a major actor in the FOSS development world to abandon a twenty-year-old battle against implementing DRM in the basic standards of core network platforms. Consumption choices appear to severely constrain the freedom of action of public-facing software development FOSS projects; interventions, if any, must be at the level of shaping demand, on the model of ethical or environmentally conscious consumption campaigns, rather than focusing solely on ethical design.

From the early days of public adoption of the Internet, there have been those who have seen decentralization primarily as a threat, empowering the nefarious, from criminals and pirates to pedophiles and terrorists to run-of-the-mill trolls and spammers. But because adaptive, flexible, loosely coupled systems were more likely to improve innovation and resilience in the face of rapid change and high uncertainty than controlled, optimized, well-behaved systems, the original Internet's design reflected a sensibility that treated stasis as far more detrimental than disruption. Unless one is willing to claim that, on balance, that assumption was wrong for the past thirty-two years, that the next thirty-two years are likely to be less rapidly changing and uncertain, or that the risks that agility and rapid innovation present vastly and reliably outweigh their benefits, it seems that the Internet's original design sensibility should continue to guide our future design choices. While defending that commitment is beyond the scope of this essay, I here outline a set of design interventions and challenges implied by present concentration trends, for those who wish to preserve the decentralizing effects of the early Internet.

Major companies and the state are the primary loci of centralizing power in contemporary society. One of the core lessons of the Internet has been that with the ap-

propriate platforms, individuals acting in peer networks can cooperate effectively without relying on the state or the market. In doing so, they create their own (however imperfect) alternative platforms for interaction, which, in turn, impose different constraints than do state-based or market-based organizations. That diversity of constraint (rather than an unattainable absence of power) allows individuals to bob and weave between different efforts – from diverse sources – to impose power on them. This both diffuses some of the centralized power and creates avenues for decentralized power.

User-owned and commons-based infrastructure are one major space of intervention. Perhaps the clearest design targets are the emerging wireless networks necessary to ubiquitous computing, including both handheld networks and the Internet of Things. For many years, proprietary spectrum allocations owned by wireless carriers – coupled with proprietary cell towers – were deemed necessary for mobile computing. It has now become clear, to the contrary, that unlicensed wireless allocations (spectrum commons) running over small-cell networks, owned by diverse organizations and individuals, are likely to be the infrastructure of first and last resort for data, with large-cell proprietary spectrum networks offering the backup for highly mobile, latency-sensitive communications.[18] The main challenge to leveraging this fact into a decentralization of power over wireless networks is to design technical and contractual systems that can permit unrelated individuals to share access to their diversely owned wireless spots. With the exception of relatively few community networks, most widespread WiFi networks are operated by companies like BT Group's system in the United Kingdom or Comcast's emerging model in the United States. Nothing technical prevents these companies' consumers from sharing their access with each other without the carrier. The constraints, instead, are contracts and social habits. One of the core design targets of any future effort to keep the Internet open, decentralized, and resistant to control is to develop technically instantiated mechanisms to achieve user-owned and -shared capacity that offers no proprietary point of control for centralizing actors.

What is true of wireless also holds for cloud storage and computing resources, though it may be more difficult to implement. Past efforts to develop distributed storage or computing include computer scientist Ian Clarke's Freenet, an early peer-to-peer data storage and communications network focused on assuring a secure system for dissidents. Oceanstore, a storage utility built atop an infrastructure of servers, and developed at the University of California, Berkeley, was a later development. Freedombox is an aspirational plug-server architecture proposed to create secure, user-owned servers that would offer much of the robustness and temporary scaling of servers provided by corporate actors, without the centralization of power. These efforts outline a critical area of open infrastructure innovation necessary to counter the centralization effects of cloud storage.

Another major design question concerns open defaults. In the case of the Android app stores explored above, Android OS phones' default settings do not permit sideloading. In WiFi devices, closed, encrypted networks are the default setting. Even though these defaults can be overridden by the user, long-term experience suggests that defaults stick. A critical target of consumer advocacy needs to be for firms that sell infrastructure and basic tools to ship them with open and secure defaults, so that user choice becomes the easy default option.

Open standards, FOSS, and law in the handheld and app-store space must also be directed to open these major control points. Deconcentrating power around the hand-

held and the app store suggest, first and foremost, efforts to develop alternatives through Web-based standards. HTML5 created the possibility of creating the look and feel of an app using an open-Web interface that need not be downloaded from an app store. As of 2015, substantial numbers of developers use HTML5 for its capacity to run across platforms, and its independence from platform-specific training and knowledge. But at this stage, it appears to sacrifice performance and optimization for generality. As long as this is true, and the rate of improvement in handheld operating systems is high, it seems unlikely that the general Web standards–based application development environment will outpace native application development. The power of the app store will remain.

An alternative would be the development of a FOSS handheld operating system (OS), such as the OS that the Mozilla Foundation is developing. As in the case of the Firefox browser, the presence of a FOSS alternative, with a strong institutional basis incorporated as a foundation dedicated to keeping the platform open, can play a role in preserving an open, decentralizing Internet. However, as the earlier discussion of DRM clarifies, that affordance is not an absolute bulwark against centralization; it is, nonetheless, a pathway to preventing additional concentration of power around the app store. If both pathways fail, it is possible that app stores will reach a point when they exercise so much control over effective access to a majority of Internet users that a legal intervention will be necessary to require app-store owners to adopt some form of nondiscrimination policy. Legal action may also be necessary to change defaults so that an app developer can initiate including itself in the app store, and the owner can only constrain access under well-specified, harm-prevention terms.

The adoption of strong, user-controlled encryption by default is one design intervention that seems both feasible and, on balance, justified. By "user-controlled," I mean encryption that provides affordances to the owner of the device on which the encryption is implemented, and constrains action on that device by others. This is by contradistinction from DRM software, which also involves end-device encryption but treats the device owner as the potential attacker, and permits some external third party (such as the copyright owner or the employer of the device owner) to use the encryption to control both uses of and access to the device. Universal strong encryption protects against both centralizing forces – primarily states and companies other than those with which the user has contracts – and decentralized sources of power, such as black hat hackers (crackers), thieves, and terrorists.

The primary opposition to adoption of universal strong encryption comes from those who suggest that the risks associated with technologically supported decentralization outweigh its benefits, and that the risks of centralization can be counterbalanced by institutional constraints on the centralizing power more flexibly and accurately than by technical barriers managed by users. The primary position of major governments is that bodies like the FBI or the NSA, properly constrained by legal oversight, will do far more good than harm if they can access any communication or device. The basic problem with this argument is that it assumes both the effectiveness of the government agencies responsible for order, and the effectiveness of the institutional controls.

As the Internet of Things blossoms, the sheer magnitude of data flows and potential points of attack becomes overwhelming to any system that seeks to read all networked information, predict events based on this data, and interdict those events. By contrast, the possibility of protecting targets locally at the individual-device level

substantially increases the cost and difficulty of harming devices and the data they store, or the processes they control. Defense will be largely imperfect, particularly against a determined and focused attack, but abuse will be more contained than with a universally less-secure system.

Moreover, the assumption that abuses by governments or companies can be adequately constrained by institutional and organizational processes is questionable at best. First, it applies, at most, to democracies with robust rule of law. For billions of Internet users in countries with weak or no rule of law, ubiquitously available strong encryption is the sole defense against abuses. Second, in democratic countries, the fifteen years since September 11, 2001, have seen persistent, repeated, and pervasive violations of human and civil rights and a persistent reluctance by authorities and courts to redress government excesses and mistakes. Multinational companies, in turn, often use jurisdictional arbitrage to escape regulation legally. The fact of the matter is that institutional systems are highly imperfect, no less so than technological systems, and only a combination of the two is likely to address the vulnerability of individuals to the diverse sources of power and coercion they face.

Future design must also take into account the resilience, redundancy, and diversity of systems resources and pathways. A central lesson of the original Internet design – its successes and failures – is that perfection is a fool's errand. Complexity is a basic condition of a connected, dynamic, open society, and with it comes persistent uncertainty and imperfection. Just as the original Internet design rejected perfectibility and optimization for openness, loose-coupling, and continuous experimentation, learning, and adaptation; so, too, must the future Internet. Any effort to finely design the environment so that it will generally permit legitimate power to flow in the legitimate direction, but constrain illegitimate power, will fail often and, sometimes, spectacularly. We need systems that are resilient, robust, and rich in redundant pathways that are open to users to achieve any given range of goals they adopt for themselves. Freedom from power, in this context, inheres in diversity of constraint; and freedom of action is maintained by bobbing and weaving between diverse efforts to impose power on the individual, rather than by following prescribed paths, such as asserting one's rights through proper channels or living on a mountaintop. The practical implication of this rather abstract statement is a simple one: design efforts need to resist calls for optimization and greater control by trusted parties if these come at the expense of open, redundant pathways and resilient capabilities.

One way of constraining power in various arenas is to create mechanisms for assuring distributed audit and accountability, rather than permission. We have auditors in government bodies and require independent auditors to certify company books; the rising call for police officers to wear body cameras so as to deter police abuse and enable redress are also (highly contested) examples of technologically instantiated audit and accountability systems. So, too, could one imagine building an effective audit and accountability system into the Internet design to enable identification and accountability of abusive power. A major concern with any such system is that it would itself create a point of centralization: in the hands of whoever controls the audit trails, or breaks into them.

It is also possible that approaches based on the blockchain could provide a useful space for developing automated audit trails. Blockchain, the technology underlying the cryptocurrency Bitcoin, is still in its infancy. But the core design characteristic may outline a solution for distributed audit trails and accountability that would avoid the

risks of reconcentration. At its core, the technology consists of three components. The first is a ledger that records all assets and transactions in a given domain. The second is encryption, which protects this ledger from tampering. And the third is distributed, redundant storage with mutual accountability such that tampering anywhere becomes evident unless it can be achieved everywhere simultaneously. This outlines an open system that would nonetheless withstand many attacks (both official and unofficial) and provide distributed users with a higher degree of confidence that abuse can be traced, documented, and ultimately fed into a system of accountability than might be possible with a more centralized and institutionalized audit system. Of course, real world accountability will require institutional and organizational adaptations; an automated audit system, decentralized or otherwise, will not be self-executing. But building an audit system with a distributed, robust architecture may offer a technical foundation around which institutions can develop.

A final proposed space for design intervention is user-owned and/or ethical governance in platforms. One of the most remarkable features of the early Internet was the emergence of working anarchies as functioning organizations with substantial social and economic impact. The IETF was the clearest example, in which an organization with practically no recognized order, functioning on self-organized, distributed, discursive arrangements independent of market, state, or other well-behaved sources of accreditation or empowerment, came to manage the core piece of global infrastructure of the late twentieth century. FOSS projects and Wikipedia followed, as the idea of self-motivated action and effective, collective work in self-governing communities matured and came to fulfill a significant part of our core utilities in networked society and economy. As these organizations matured, they began to develop hybrid approaches, mixing formal nonprofit incorporation with internal meritocratic, nonhierarchical structures (such as the W3C, the Apache Foundation, and the Mozilla Foundation), or independent community structures, alongside and of superior legitimate power than the formal foundation set up alongside them (Wikimedia Foundation and the Wikipedia community). As we look ahead toward the design of the future Internet, many challenges will appear to require structured organizational responses, like state-based agency intervention or market-based, proprietary companies. What the past twenty years of self-organized communities suggest is that peer production and social self-organization offer a diverse and rich design space for solving collective action problems and implementing organizational effectiveness without necessarily falling into the trap of state or market, and without simply permitting the emergence of unaccountable oligarchies instead.

When the Internet was first designed, few knew about it, and fewer understood its significance. The major design decisions were made in a power vacuum. By now, everyone knows that Internet-design decisions will affect political, economic, institutional, social, and cultural arrangements, and decisions that will influence the next quarter-century are all being influenced themselves by sustained efforts of diverse parties that stand to benefit from them.

Much virtual ink has been spilled on democracy, innovation, privacy, and cyberhacking, which all address the fundamental problem of power. In all these more familiar framings, how the Internet enables or disables some people to influence the perceptions, beliefs, and behaviors, as well as the outcomes and configurations that other people hold and inhabit, is at stake. In the second half of the twentieth centu-

ry, core values of individual autonomy and self-authorship, creativity and ingenuity, community cooperation, and collective self-governance were all associated with representative democracy; civil rights; the rule of law in property, contracts, and the state; coordination through prices in markets; and stable social institutions, like the family, church, union, and civic association. In the past quarter-century, looser associations have become effective, while these more traditional institutions continued to offer some degrees of freedom and effective action, but also became sources of constraint vis-a-vis the new forms of action and association.

As we struggle with diverse design choices, it is important to recognize the substantial emancipatory and creative power of the open and loosely coupled action systems that the early Internet enabled and empowered. Their force in supporting creativity, autonomy, and chosen association is often linked with relatively weaker governability and less-focused capacity to express a coherent voice. While we have had examples of successful collective action by distributed, Internet-enabled forces over the past few years, the steady grind of policy-making and standards-setting mean that the values of a genuinely open Internet that diffuses and decentralizes power are often underrepresented where the future of power is designed and implemented. Thus, it falls to those primarily in the relatively independent domain of academia to pursue these values and insist on diagnosing design choices in terms of their effects on the distribution of power, as well as to develop and advocate design options that will preserve the possibility of decentralized, autonomous, and organically chosen collective action. Our alternative would be transmitting the power of those organizations that have the wherewithal to sit at every table, and in every conference room, to assure their own interests in the design of our future.

ENDNOTES

1 Tyler Lopez, "How Did Ukraine's Government Text Threats to Kiev's EuroMaidan Protesters?" *Slate*, January 24, 2014, http://www.slate.com/blogs/future_tense/2014/01/24/ukraine_texting_euromaidan_protesters_kiev_demonstrators_receive_threats.html.

2 David Clark, "A Cloudy Crystal Ball – Visions of the Future," in *Proceedings of the Twenty-Fourth Internet Engineering Task Force*, ed. Megan Davies, Cynthia Clark, and Debra Legare (Cambridge: Massachusetts Institute of Technology, 1992), 539 – 545, http://ietf.org/proceedings/prior29/IETF24.pdf.

3 comScore, "The U.S. Mobile App Report," (Reston, Va: comScore, 2014), http://cra.org/wp-content/uploads/2015/02/The-US-Mobile-App-Report.pdf.

4 Jonathan Zittrain, *The Future of the Internet and How to Stop It* (New Haven, Conn.: Yale University Press, 2008).

5 Jon Brodkin, "Netflix Performance on Verizon and Comcast Has Been Dropping for Months," Ars Technica, February 10, 2014, http://arstechnica.com/information-technology/2014/02/netflix-performance-on-verizon-and-comcast-has-been-dropping-for-months/.

6 Measurement Lab Consortium, *ISP Interconnection and Its Impact on Consumer Internet Performance*, October 28, 2014, http://www.measurementlab.net/static/observatory/M-Lab_Interconnection_Study_US.pdf; and MIT Information Policy Project, in collaboration with the UCSD Cooperative Association for Internet Data Analysis, *Measuring Internet Congestion, A Preliminary Report* (Cambridge: Massachusetts Institute of Technology, 2014), https://ipp.mit.edu/sites/default/files/documents/Congestion-handout-final.pdf.

7 Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock, "Experimental Evidence of Massive-Scale Emotional Contagion Through Social Networks," *Proceedings of the National Academy of Sciences* 111 (29) (2014): 8788 – 8790.

8 Robert M. Bond, Christopher J. Fariss, Jason J. Jones, et al., "A 61-Million-Person Experiment in Social Influence and Political Mobilization," *Nature* 489 (7415) (2012): 295 – 298.

9 Zeynep Tufekci, "Engineering the Public: Big Data, Surveillance, and Computational Politics," *First Monday* 19 (7) (2014), http://firstmonday.org/ojs/index.php/fm/article/view/4901/4097; and Jonathan Zittrain, "Facebook Could Decide an Election Without Anyone Ever Finding Out," *The New Republic*, June 1, 2014, http://www.newrepublic.com/article/117878/information-fiduciary-solution-facebook-digital-gerrymandering.

10 Shoshana Zuboff, "Big Other: Surveillance Capitalism and the Prospects of Information Civilization," *Journal of Information Technology* 30 (1) (2015): 775 – 789.

11 President's Council of Advisors on Science and Technology, *Big Data and Privacy: A Technological Perspective*, (Washington, D.C.: Executive Office of the President, May 2014), xii, https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf.

12 Ibid., xi.

13 Cory Doctorow, "We Are Huxleying Ourselves Into the Full Orwell," Mostly Signs, Some Portents, January 9, 2014, http://mostlysignssomeportents.tumblr.com/post/72759474218/we-are-huxleying-ourselves-into-the-full-orwell.

14 Proportions calculated by author from Nielsen, *The Total Audience Report* (New York: Nielsen Company, December 2014), http://www.nielsen.com/content/dam/corporate/us/en/reports-downloads/2014%20Reports/total-audience-report-december-2014.pdf.

15 See Sandvine, *Global Internet Phenomena: Asia-Pacific & Europe* (Waterloo, Ontario: Sandvine Incorporated, 2015) as discussed in Adam Epstein, "Netflix Now Accounts for 35% of Bandwidth in the U.S. and Canada," *Quartz*, November 20, 2014, http://qz.com/299989/netflix-now-accounts-for-35-of-bandwidth-usage-in-the-us-and-canada/.

16 Nielsen, *The Digital Consumer* (New York, The Nielsen Company, February 2014), http://www.nielsen.com/content/dam/corporate/us/en/reports-downloads/2014%20Reports/the-digital-consumer-report-feb-2014.pdf.

17 Kristen Purcell, "Online Video 2013" (Washington, D.C.: Pew Research Center, October 10, 2013), http://www.pewinternet.org/2013/10/10/online-video-2013/

18 President's Council of Advisors on Science and Technology, *Big Data and Privacy: A Technological Perspective*; and Yochai Benkler, "Open Wireless vs. Licensed Spectrum: Evidence from Market Adoption," *Harvard Journal of Law & Technology* 26 (1) (2012): 69 – 163.

Edge Networks & Devices for the Internet of Things

Peter T. Kirstein

Abstract: This paper considers how existing concepts underlying the development of the Internet are being strained in the emerging Internet of Things (IoT). It also explores how the well-known and tried Domain Name Service concepts, which map hierarchic names to addresses, can be extended for the IoT. The extension greatly broadens the concept of name/address mapping to digital objects with identifier/attribute database mapping for physical objects, applications, and data. Finally, this essay discusses the properties of the identifier management systems, showing how scalability, security, and flexibility can be supported in the IoT.

PETER T. KIRSTEIN, a Foreign Honorary Member of the American Academy since 2002, is Professor of Computer Systems at University College London. He is the author of *Space-Charge Flow* (1967) and has recently published articles in *Sensors, Information Systems,* and *International Journal of Informatics Society*.

The initial aim of the Internet was to develop a system that would allow computers to connect together, irrespective of their location or individual method of connection. That system grew to connect the several billion systems in use today. The Internet is now becoming the *Internet of Things* (IoT), embracing the hundreds of billions (or trillions) of digital devices that can sense or activate aspects of our lives. The IoT is still in its infancy: the state of networks and variety of equipment types in the IoT today is comparable to that at the onset of the Internet, from 1975 – 1980. In this essay, we explore some of the theory behind the design of the Internet, and consider the ways in which the needs of the IoT fundamentally differ. At the same time, we will examine similarities between the development and growth of the Internet and the IoT. Of course, Internet protocols (IP) have developed hugely since the Internet's youth. We will not consider the core of the Internet in this paper, but will focus on what new demands the IoT may make on that core.

Even thirty-five years ago, it was clearly important to link together the many network deployments of different architectures. But there was a question whether to choose to *adapt* between network types or wait for universal *adoption* of the same type. Adoption eventually won out, though it took fifteen years.

doi:10.1162/DAED_a_00363

The design of the early Internet assumed that any compatible networks would use the Internet protocol IPv4 at the network level. Because the ARPANET (Advanced Research Projects Agency Network) could address only 256 computers with permanent identities, computer scientists considered the increase to four billion machines, as allowed by IPv4, as more than would ever be needed. Within about fifteen years, however, it was apparent that even this would be inadequate. Various short-term measures were taken to allow the mechanisms to cope, including introducing private addresses that allowed address space to be reused, albeit with a loss of flexibility, which made it very difficult for a device to have more than one address. While these measures allowed the Internet to continue to grow without adopting a new universal network type, it had become clear that the existing structure could not cope with the huge numbers envisioned even without the advent of the IoT.

In order to plan for the inevitable address crunch, the Internet community decided to specify a new protocol: IPv6. This resolved the addressing problem and fixed a number of other shortcomings: allowing multiple addresses for a single interface, group operations, limiting the scopes of addresses, and improvements in mobility support, among other gains. Although aids have encouraged transition, moving all new customers, let alone existing customers, from IPv4 to IPv6 has proven to be a long and difficult process. Yet this is now occurring on an increasing scale, particularly for newer applications and in contexts in which customers are running out of IPv4 address space. While there are various attempts to design completely different network architectures and components, IPv6 will prevail.

To cope with the relatively large number of computers that the initial Internet intended to connect, it was necessary to define some human-friendly directory of computers; hence, the *Domain Name Service* (DNS) was defined to connect user-friendly names to Internet addresses.[1] Engineers deployed a scalable architecture, which has lasted through the introduction of IPv6. The system is hierarchic, meaning the owner of the domain has almost complete freedom at any level to allocate address ranges – indicated by a "." – to the domains below it. For kirstein.cs.ucl.ac.uk, for example, the .uk domain is allocated a large block of addresses: it has jurisdiction over a number of domains including commercial (.com), nonprofit (.org), government (.gov), and academic (.ac). The registered owner of .ac allocates from his range of addresses a set to each university (such as University College London: .ucl), which in turns allocates a range of its addresses to each department (such as computer science: .cs). The Internet assumes that all end points obey the Internet protocols (IPv4 or IPv6), and so the only value that has to be returned from a query to the DNS is the IP address. There is very limited security: the owner of a domain like cs.ucl.ac.uk enters certain security features to ensure that only authorized entities may insert name/address pairs. The implementation of the DNS has evolved over the last thirty-five years; it is highly distributed and many parts are replicated for resilience. The lowest levels are often near the end systems, and the number of entries on a particular platform is kept reasonably low to maintain performance. The system has continued to perform with the few billion names it now contains. Individual user processes often cache the information of often-used end devices to minimize further access delays.

In the original Internet, it was generally assumed that each interface to a computer was attached to a unique network and had a unique name. Thus, the name/address could be unique. More recently, with the advent of both wireless networks and IPv6, this uniqueness has been put into doubt.

The same wireless interface can be seen by a number of different overlapping networks, and the same interface can have a number of different names (as seen by different application operators).

Over the years, the DNS was modified in three important respects: the capability for adding descriptions of services (DNS-RR),[2] the ability to search for services (DNS-SD),[3] and the capability to authenticate DNS entries (DNSSEC).[4] While parts of the RR (Resource Records) can be encrypted, access to all attributes of the entries has remained open. All entries to the DNS are assumed to obey the IP suite, though the RR gives some information on how services can be accessed.

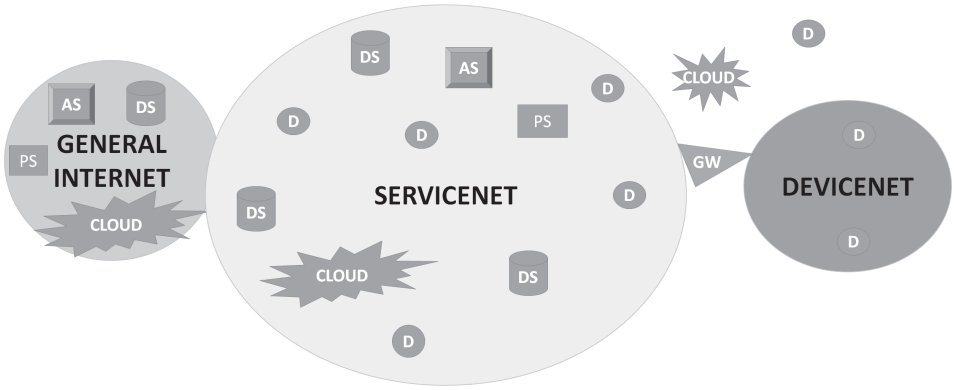
With the move toward the IoT, many of the fundamental assumptions of the Internet are overturned. It was initially assumed that we would only communicate with computers at the edge of the networks that make up the Internet. More recently, the scope of edge devices has broadened, including personal computers, telephones, printers, and SatNavs (satellite navigation). However, the digital controllers in these devices all obey the Internet protocols. If there are changes in these protocols, we can assume that most of the devices will evolve so that their successors can incorporate the changes. With the IoT, this may be the case, but some areas, such as building automation, bridges, and ships, may have much slower rates of change. There are already many standards for automation systems, which often now use Internet interfaces, though these usually just allow the same procedures to be carried out remotely as were previously performed locally. In these systems, the whole concept of network, edge device, and network technology is much broader.

Figure 1 illustrates a particular application (using IP technology) running over a specific network we call the *ServiceNet*. The application will be connected to the Internet, but also to many different devices (D). Some devices may be IP-enabled, some not. The latter are shown connected through a gateway (GW) to a network called the *DeviceNet* (T). The ServiceNet is related to a physical deployment of devices, gateways, data storage elements (DS), process servers (PS), and application servers (AS). Some of these may be free-standing; others may be on processors in a computing cloud (CLOUD). The various processing elements may be directly related to the deployment; then they are shown as being located on the ServiceNet. They may be much more remote entities, shown in Figure 1 as on the general Internet. It is important to understand that while there may be a large number of real objects deployed in the IoT, the ServiceNet is a virtual network of the subset used in a particular application.

Figure 2 highlights the distinction between a *DeploymentNet* (DNET) and a ServiceNet. There may be a number of different deployments, each characterized by a single owner and database. An example might be individual smart buildings, in which the DNET refers to the entities that are a subset of those within that building. An application might refer just to one such deployment: for instance, all the temperature gauges or lights in that building. It might, however, refer to entities in several buildings: such as the set of fire alarms or electricity meters on the whole street. This is indicated in Figure 2 by calling the ServiceNet an *Application-ServiceNet* (APP-SERVICENET). The network connecting all the devices in specific applications is thus a *virtual network*. Different applications may be concerned with different subsets of devices in the different deployments.

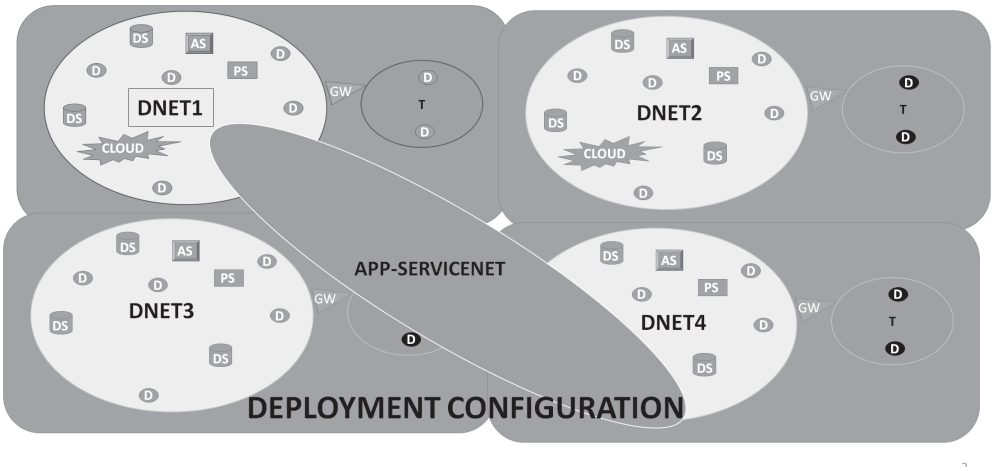
This is indicated in Figure 3 by the different APP-SERVICENETs shown. In light of this, one view of the deployments is the *deployment configuration*: the collection of all the physical devices deployed. Normally, there would not be a single database or description illustrating this collection; rather,

Figure 1
Basic Network Diagram

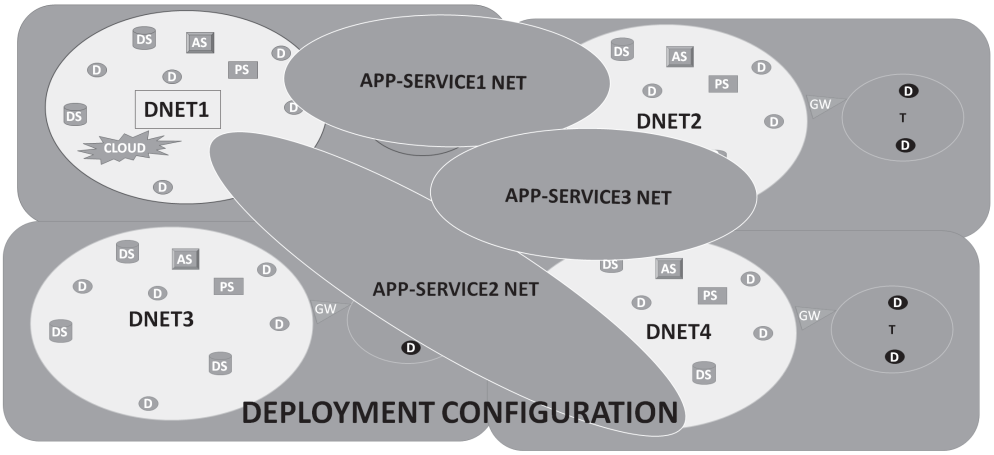


Relevant IP-enabled devices are located on the ServiceNet; those that are not so enabled are on the DeviceNet via a gateway. The servers are on the ServiceNet or general Internet. Key: devices (D); gateway (GW); data storage elements (DS); process servers (PS); application servers (AS); computing cloud (CLOUD). Source: Figure prepared by the author.

Figure 2
Single Applications in Complex Configuration

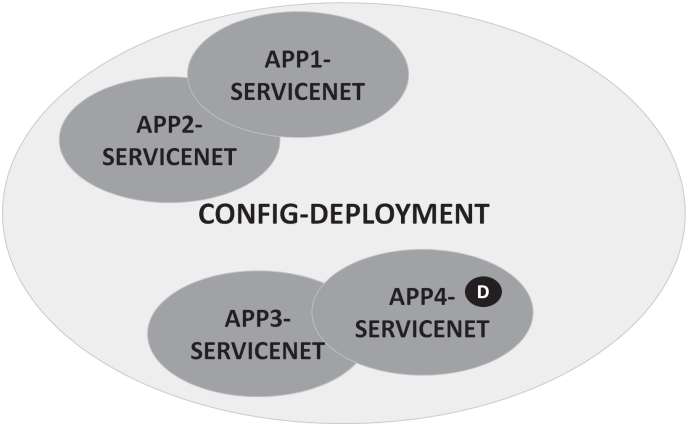


Deployments belonging to different entities are shown on different DeploymentNets. An Application-ServiceNet will use a subset of devices that may be on several DeploymentNets. Key: devices (D); gateway (GW); data storage elements (DS); process servers (PS); application servers (AS); computing cloud (CLOUD); DeploymentNet (DNET); DeviceNet (T); ServiceNet/Application-ServiceNet (APP-SERVICENET). Source: Figure prepared by the author.



Several different applications can use the same DeploymentNet configuration. Sometimes several applications can use the same devices in different ways. Key: devices (D); gateway (GW); data storage elements (DS); process servers (PS); application servers (AS); computing cloud (CLOUD); DeploymentNet (DNET); DeviceNet (T); ServiceNet/Application-ServiceNet (APP-SERVICENET). Source: Figure prepared by the author.

Figure 4
Multiple Applications on Given Deployment



An alternate representation of Figure 3, showing a general deployment and several applications using the same or different devices that use subsets of that deployment. Key: devices (D); ServiceNet/Application-ServiceNet (APP-SERVICENET); configuration deployment (CONFIG-DEPLOYMENT). Source: Figure prepared by the author.

there will be separate deployments and databases referring to different management units, as is shown clearly in Figure 2.

If we now consider how applications may interact with physical deployments, the picture becomes even more confusing. In fact, there is a third type of entity in the IoT: namely, *data*. In some forms of deployment, sensors produce data as a result of interactions from an application. In others, the deployment is such that the data are produced spontaneously and continually. Each block of data may then contain certain *metadata* about the data, which may contain attributes that indicate the source of the data, the time it was produced, the location of the producer, both authentication information to ensure it was produced as stated and security information to indicate how it may be used, and even ownership information. The task of dealing with these superpositions of deployments and applications is too complex in real space, particularly if we try to depict the use of the same device or data by different deployments or applications. It is much more effective to work in cyberspace, provided that we can show clearly how to relate the cyberspace to the physical space that we think we better understand (see Figure 3).

The situation illustrated in Figure 3 may be generalized, as shown in Figure 4. Here the set of all devices and subsystems in a particular environment is called a configuration deployment (CONFIG-DEPLOYMENT). The different applications form specific ServiceNets. Sometimes the applications are quite distinct and use different devices; sometimes, however, they overlap by using some devices in several applications. For example, the collection of air conditioners in a building might be managed in one application. The set of all air conditioners of a particular type in a city may be managed for servicing purposes by another entity. Provided the details of the deployments are known, it is possible to plan whether a particular application is feasible and useful. Moreover, by slightly extending the deployment, whole new applications might be achievable at marginal cost.

It is clear we need a more holistic description of devices, deployments, applications, and data than has been attempted so far. The key to this expansive description is to consider not specific devices or networks, but just *digital objects* (DOs). We may work with DOs, their identifiers, and their attributes and denote this as working in cyberspace. Digital objects are a much broader concept than physical objects. The different perspectives of the same deployment described above become simpler here: it is now represented by different DOs. A DO is defined by its identifier; this is a string of bits that represent the name of the object, which usually has some hierarchic structure. The identifier is associated with a set of attributes that describe its properties. Thus, the air conditioning unit as seen by the building's operator can have an entirely different identifier from that seen by the service engineer. But DOs need not be associated only with physical objects; they can also be associated with processes and data objects. It all depends on the attributes associated with the DO. All the discussion on DOs, particularly about the properties desired, is heavily based on the work of electrical engineer Bob Kahn and the Corporation for National Research Initiatives (CNRI), who have implemented the Handle System.

With every DO, one must associate an identifier: a handle by which we can refer to, and work with, the DO. The most general way to describe the DO's properties is through a set of attributes, each described by a *type/value pair*. The first describes the nature of the attribute, the latter describes some value. Of course, there must be a description elsewhere of what is meant by that *type* and how it is represented by the *value*. The identifier may be associated also with

metadata that describe how the attributes may be accessed – and possibly with its owner. To describe a large number of DOs, there should be some form of registry of types, which may even reveal the way the values of DOs will be described. The attributes should be stored in a searchable database, allowing applications to ascertain which DOs they might wish to use. Thus, we can make DOs useful by associating them with an identifier management system (IMS) comprising three parts: an identifier resolution system (IRS), an identifier attribute store (IAS), and an identifier type registry (ITR). For each physical object deployed, a DO is defined. Of course, different stakeholders may have different views of the physical object – as with the air conditioner – thus, it may be associated with several identifiers.

It is possible, and useful, to define applications and data elements as DOs. This allows the metadata either to explain how the DO can be accessed or to reveal attributes of the DO itself. Thus, applications may be stored generically, including a template. A specific application can thus be defined in terms of inserting the details of a particular deployment into the template of the generic application. Similarly, data may be stored as a DO with the appropriate attributes; applications may even simply reprocess the data for a new purpose, providing, of course, that it is authorized to do so. Just as an application might use a subset of the physical deployment, it may use a subset of the relevant stored data for its purposes.

Much of the activity involving DOs can be processed via normal computer clouds; sometimes, however, fast processing is required. In that case, one requires substantially more local processing, o r computer clouds with definable quality of service (QoS) standards of performance.

The IoT has Internet in its makeup, so our analysis should use as many of the tried properties of the Internet as possible. The DNS employs a hierarchical structure, and its implementation architecture has shown that it can be distributed at will; the properties of control of the IoT DO identifiers fulfill exactly the same need as does the DNS. While the DNS was deliberately open for all entities to access, this is not necessarily desirable in the identifier attribute store. While universal accessibility is appealing, the deployment owner may require that access to some of the attributes be limited. Thus, while the original Internet deliberately introduced a minimum of security in the DNS, the IoT would benefit from constraining operations on the IAS. Indeed, it would be desirable to constrain the authorization to *create*, *delete*, *modify*, or *access* both identifiers and attributes, achievable through association with appropriate metadata. We may go further still and ensure that any values of attributes transmitted from the IAS must be encrypted.

There are already many large component databases whose data could, if made available, describe directly the DOs of digital devices; for this reason, it is useful to define one attribute as being the ID in any other such database of interest. We have already given examples of how some DOs may themselves represent complex systems; thus, having another *attribute type* allows the system to be recursive. To tie the cyberspace representation to the physical world, we must have another type of IP address (or name). Provided IPv6 is used, there is no reason why the same physical object cannot be represented by DOs with different identifiers and IP addresses.

With the Internet, the planning of deployments has not been a major part of the network engineer's work. With the IoT, the configuration process and the application design and implementation are central concerns. Further, during the physical deployment phase, the population of the IAS is of vital importance. Most physical de-

ployments follow some domain-specific procedures. For example, in the construction and functioning of a smart building, the architect, installation engineer, building supervisor, security officer, and service engineer each have unique roles. Normally, drawings and specifications are produced as part of a business process; processing algorithmically the models of the physical systems into DO form will be a major aid to implementing large-scale deployments in the future.

This will require the development of tools that can populate the IAS algorithmically based on the data already extant in different domains. This process requires the provision of security tokens; thus, every physical entity that may need to be secured on actuation will need one or more security tokens and an associated list of authorized entities. Similarly, the data of every sensor that provides information that may require authentication should be signed by the authentication token associated with the device. During the setup phases of configurations, attributes required for authentication or actuation should be put in the IAS. In some cases, such as when asymmetric encryption is used, the entry is not sensitive. In others, it is critical that it be stored only in an encrypted form. During operations, the IMS may be used to control proxy security operations for devices too constrained to do them locally. In order for these tools to work well, it is likely preferable to be able to define some *templates* on what attributes are permitted and needed by the entity using the tool. Note that any time there is a change to the configuration – for instance, if a sensor is replaced – it may be necessary to update the IMS; if only to provide a new security token.

For many situations, these physical deployments will be constant. Thus, for example, each building, traffic light system, or surveillance system connected to the IoT may have different physical models that need to be processed to populate the IMS. An application will often deal with a subset of the whole configuration. This might be termed a *virtual deployment*. The individual deployments may belong to different entities. Because there may need to be a negotiation regarding the terms by which an application can use parts of a physical deployment, one part of the metadata associated with an identifier may have to be its ownership. In a typical life cycle, an application will be designed, implemented, deployed, and put into operation. Having determined the usage rights for a deployment, the information in the IAS is used to define the application in the design and implementation phases. Some devices in the IoT may require special processes to access them; this will be specified in attributes stored in the IMS.

It is important to note that during the phases of designing, implementing, and deploying physical infrastructure, data is put into the IAS as part of the deployment process. During the design and implementation of applications, data in the IAS is used to define the virtual configuration appropriate for the application.

Some massive applications do not require access to physical deployments and their related applications. It is adequate to access only the data previously stored. Indeed, this property is at the heart of many of the current generation of large start-up enterprises like Google and Facebook. Their data are produced from a different set of applications and deployments; it is only the authorization to use and deep-mine the data that their applications require.

In the IoT, certain compound operations can be very convenient, such as *reading all sensors on a floor* or *notifying all cars in a particular location of a nearby accident*. Of course, it is possible to define such operations in an application; however, it may make both the design of the application and its opera-

tion simpler if the relevant operations can be carried out in physical space. Similarly, it would be convenient for applications to use network addresses located in the address space of the owner of the application. As explained above, we can associate physical space with cyberspace by defining one attribute of an identifier to be its IP address. If the ServiceNet shown in Figure 1 is in IPv6, then both features are supported at the network level. The group operations can be supported by *multicast*, allowing operations at the network level to be performed on a group of objects. And in this form of network, there is no problem associating different IPv6 addresses with the same object in use in different applications. Neither of these functions are essential, but they certainly ease application design and operation.

There are many extant identity management systems; it is unlikely that they will all adopt the same implementation in the near future. Electronic components in particular, but increasingly also types of subsystems – lifts, cars, automation subsystems – will be stored in an identifier database complete with all their properties. Ideally, another type of attribute is the identity of a given subsystem in a different database.

Even the Internet has relied on a consistent management structure that defines protocols, allocates address space, and specifies security features. While there have been political concerns that international governance bodies such as the Internet Assigned Number Authority (IANA), the Internet Advisory Board (IAB), and the Internet Engineering Task Force (IETF) have not been appointed in the conventional manner, they have nevertheless functioned well. In the context of the IoT, even more governance is likely to be required. It is probable that when the identifier systems outlined here are globally accepted, there will be entities in each application domain that assist in the governance of, at least, the identifier space, and likely also of the attribute types that are standardized for the domain. In this realm, other bodies will be concerned with standardization across domains.

Security is the great challenge posed to those working toward stable governance, authorization, and user responsibility. On the one hand, those responsible for specific installations may need to organize their own security trust chains; on the other hand, the chains may need to be regulated by an official third party. Clearly, in terms of access to the data objects in the IAS, we are moving toward the general considerations of privacy of data, ownership of data, and permitted usage. Here we stray well beyond the past and future of the Internet. However, the provision of these broad classes of DOs inevitably leads to very difficult cases of who is entitled to what access to the IAS and under which conditions.

In retrospect, the concept of the Internet was simple compared to the Internet of Things. At the time, it seemed a daunting task to persuade less than a dozen major suppliers to change completely their protocols for connecting computers together. But it was successful because the concept was so straightforward. Of course, the Internet evolved to deal with issues of scale, heterogeneity, and performance; but the foundational concepts remained relatively stable. Three early adjuncts to the basic Internet protocols were vital: keeping heterogeneity on the edge of the network, restricting security to the edges, and setting up a highly distributed system for name/address mapping. For the IoT, many more large industrial and political players must be persuaded to adopt a common approach. Moreover, the number of edge devices in use in the IoT are many orders of magnitude greater, the governance more challenging, and trustable security more vital than with the Internet. However, the experience gained through the introduction and deployment

of the Internet gives us a much clearer indication of what is required in advancing the IoT.

The way in which the deployment of physical devices is almost orthogonal to the development and deployment of applications provides a clue as to how to proceed. The imperative of being able to scale to much larger numbers of devices, while keeping the size of individual deployment authorities and applications operators manageable, gives a second. How the complex nature of trustable authentication and authorization must be provided for edge devices that have limited computing power and memory capacity is a third. Finally, the need to be sufficiently flexible to allow different communities to adopt myriad ways of working is inevitable. The third orthogonal category of DOs fits naturally into the same basic technology; it is clearly another natural aspect of the IoT. While fundamental to the benefits, and dangers, of the IoT, it leads to whole new deployments, uses and reuses, security and privacy concerns, responsibility and liability, and domains of regulation and control.

The above considerations make it important to work conceptually (in cyberspace) as much as possible. This is particularly so in the case of physical deployments and maintenance. All maintenance of physical devices can be recorded in cyberspace, where the authentication and authorization attributes can also be maintained. This allows applications to be developed on virtual deployments, or even using existing data, which are a subset of the physical deployments and/or data derived from cyberspace databases. The scalability, with manageable subsets, can be assured by adopting the structure of the domain name service, the power of the identifier management system, and the flexibility of allowing attributes to refer to an arbitrary set of other identifier systems. Finally, the growth of cloud computing allows most of the cyberspace work to be carried out in the computer cloud, while operational concerns are carried out in local services, which probably also adopt local clouds with specifiable characteristics. Authorities have stressed the importance of deploying ServiceNets based on the newer Internet protocol IPv6 because of its larger address space capacity. We go further, having considered how use of IPv6 gives important advantages in multistakeholder use of shared interfaces and in enabling the group operations common in the IoT.

ENDNOTES

1 P. Mockapetris, "Domain Names – Implementation and Specification," *IETF RFC 1035* (November 1987), https://www.ietf.org/rfc/rfc1035.txt.

2 The TCP/IP Guide, "DNS Message Resource Record Field Formats," http://www.tcpipguide.com/free/t_DNSMessageResourceRecordFieldFormats.htm.

3 S. Cheshire and M. Krochmal, "DNS-Based Service Discovery," *IETF RFC 6763* (February 2013), https://tools.ietf.org/html/rfc6763.

4 R. Arends, R. Austein, M. Larson, D. Massey, and S. Rose, "Protocol Modifications for the DNS Security Extensions," *IETF RFC 4035* (March 2005), https://tools.ietf.org/html/rfc4035.

Reassembling Our Digital Selves

Deborah Estrin & Ari Juels

Abstract: Digital applications and tools that capture and analyze consumer behaviors are proliferating at a bewildering rate. Analysis of data from large numbers of consumers is transforming advertising, generating new revenue streams for mobile apps, and leading to new discoveries in health care. In this paper, we consider a complementary perspective: the utility of these implicitly generated data streams to the consumer.

Our premise is that people can unlock immense personal value by reassembling their digital traces, or small data, into a coherent and actionable view of well-being, social connections, and productivity. The utility of reassembling the self arises in diverse contexts, from wellness to content-recommendation systems. Without design attention to the unique characteristics of small data, however, the image that these data provide to individual users will be, at best, like a cubist portrait: a fragmented picture of the self.

Management of small data presents fundamental design questions regarding the "who, what, and where" of access rights and responsibilities. The blend of competing and cooperating entities handling small data breaks down distinctions such as that between shared and private, and renders questions like *whose data are they?* hard to answer. Conceptual boundaries blur further as data increase in sensitivity and become "activated," such as when personal apps process and fuse longitudinal data streams to drive context-rich personalization algorithms on the consumer's behalf.

We explore this confusing landscape by drawing attention to three critical design objectives: programmatic access to the digital traces that make up small data, activation of small data for personal applications, and creating privacy and accountability measures for the apps and services consuming small data. We point out the limitations of existing perspectives on both data ownership and control, and on privacy mechanisms, such as sanitization and en-

DEBORAH ESTRIN, a Fellow of the American Academy since 2007, is Professor of Computer Science at Cornell Tech and Professor of Healthcare Policy and Research at Weill Cornell Medical College.

ARI JUELS is Professor at the Jacobs-Technion Cornell Institute at Cornell Tech.

*(*See endnotes for complete contributor biographies.)*

doi:10.1162/DAED_a_00364

cryption. Rather than attempting to provide answers, we pose key questions that should inform new system designs.

The term *big data* expresses the potential of extracting meaning and value using data sets covering large numbers of people, or a large *n*. Big data's humbler counterpart, *small data*, promises to be equally transformative, allowing individual users to harness the data they generate through their own use of online and mobile services; or, in other terms, when $n = 1$.[1]

The explosion of data about individuals is no secret. Personal data sources include: continuous activity and location data sourced from mobile devices and wearables; URL and click data from online searches; text and voice data from social and personal communications (emails and posts, texts and tweets); photographs both taken and viewed; entertainment preferences and consumption; daily purchases made online and offline; personal records from digital education files and medical systems; transportation preferences and patterns; and emerging sources such as genomic data, implanted medical devices, and wearables.

In this paper, we discuss the promise and problems associated with small data. We use the term *small data* to refer to the digital traces produced by an individual in the course of her daily activities, which she can use not to understand general trends across a population, but to understand herself. All of the benefits of small data require a reassembly of the self, a partial to comprehensive drawing together of diverse small data sources pertaining to the individual. While there arise many technical challenges related to data standards, storage, and computation, we focus on the issues in greatest need of architectural attention: *access*, *activation*, *privacy*, and *accountability*.

When should a person have programmatic access to the digital traces he generates, along with the capability to activate these data through applications and services of his choosing? Several examples highlight how small data, as a complement to big data, promises powerful new insights and opportunities:

1) *Custom small-data analytics*. Consider a mobile health application that guides a patient through preparation and recovery from hip surgery. Someday, such an app could analyze her daily walking patterns, provide predictive analytics on her recovery time, and engage her in the physical therapy regimen that best matches her unique medical situation. Such approaches are expanding beyond their origins in the quantified-self movement into broad-based health management practices.[2] The perspective of small data, rather than big data, will provide not only global insights that lead to new therapies, but personalization of these therapies to the patient, time, and place.

2) *Rich user-modeling to facilitate social services*. The quantified self has a natural counterpart in the quantified student. For example, a teacher or tutor could gain great insight from a synthesis of individual students' detailed analytics, as captured by patterns in their online consumption of lectures and readings, or in their online input during homework exercises and examinations. Similar functionality could enrich relationships between mentors and mentees, coaches and clients, and provide crucial support to those whose job it is to safeguard the well-being of teens in the foster system.

3) *Service and product personalization*. Rich user-modeling is equally relevant to service personalization, recommendation systems, and advertising. Popular online platforms like Amazon and Netflix and sharing-economy services like Uber and Airbnb, are largely informed by a very narrow set of data available to them, either directly or through third-party acquisition. Imagine the immersive recommendation systems

that could be built by drawing on users' *full* suites of data, from online retail and service transactions to location and mobile communication data.

This direction could continue to be pursued strictly as a big data play to sell more products and services to targeted customers, such that utility is measured in terms of sales figures. However, we have already seen signs of customer pushback against the perceived "creepiness" of platforms mining personal data to boost sales. If individuals can, instead, demonstrably benefit from personalization on their behalf – in other words, if *utility is instead shown in terms of small data benefiting the individual* – then "getting it right" can advance whole industries beyond contention with consumers.

(4) *Enriching the arc of individual-to-community knowledge*. Individuals share data with communities to accumulate shared knowledge and a collection of experiences. Small data streams, contributed by individual users could, for instance, amplify the great success of manual data entry for sites such as PatientsLikeMe and Inspire, which help patients and caregivers understand and navigate the choices and challenges of specific medical conditions.[3] The small data perspective also points to a path for this collective knowledge to return to the individual in the form of moment-to-moment guidance. Knowledge and predictions about matters from food allergies to triggers of seizures can be mapped continuously onto an individual's small data from a bank of collective experience.

These potential benefits are uncontroversial. But controversy arises and design focus is most needed when we consider an individual's access to her own small data. In order to realize the benefits inherent in the above examples, a consumer needs to have access to her own digital traces, and also needs to be able to activate them, such as by unlocking them in one piece of software and making them available for use in another. It might seem self-evident that this combination of access and activation of one's own data is an imperative, even a universal right. But it is not.

Do you have an irrevocable right to your own physiological data? It is hard to imagine an answer other than yes. But most fitness tracking devices and mobile apps do not give users direct access to raw data on their physiological measures, such as number of steps taken, skin temperature, body weight, speed of food intake, and heart activity. Instead, users must upload the data to a device- or software-maker's service for analysis and display. Users often cannot download or export this raw data because the makers of fitness devices and apps frequently rely on business models that exploit control of their users' data and outline terms of use that claim broad rights to user-generated data.[4]

Tensions around the rights of the individual to her physiological data are not new. But in the past these concerns primarily affected the small segment of the population with implantable medical devices, such as pacemakers and insulin pumps.[5] Now, these issues of physiological data use and ownership impact every user of a mobile phone, smartwatch, or fitness-device.

One complication is the fact that the physiological data recorded by apps are created not just by an individual, but in collaboration with an app. The mobile app that records your footstep is, in fact, a collaboration between your body, which produces the motion, and the accelerometers in your mobile device, which detect it. Their outputs are then translated by the app (or a cloud service) into human-consumable data, like pedometer readings. Moreover, the model that translates the data most likely benefits from data from other users, further complicating the issue!

Joint creation is a widespread feature of small data. If a user interacts with a service provider's content – such as when buying a

video online or posting a comment on the provider's site – ownership and control of the resulting data, transaction, or text can be a complicated matter. Activities like taking group photos, videoconferencing, and gathering nutrition data on shared meals all result in the joint creation of small data.

As the amount, variety, and multiplicity of stakeholders in small data balloon, the questions of rights and control become increasingly complicated.

If people are going to have ready access to sensitive information about themselves, what platforms and methods will support the privacy and accountability needed for apps and services fueled by these data? Small data can furnish powerful insights, for good and ill. So it is essential that the technical community start to develop mechanisms and build products that allow users to access and activate their small data, while protecting them from abuses in digital and commercial ecosystems far too complex for them to reason through, let alone manage.

Today's exploration of privacy violations foreshadows tomorrow's challenges in small data protection. Consider this example: *On July 8 at 11:20 a.m., Olivia hailed a taxi on Varick Street in the West Village in Manhattan. An eleven-minute ride brought her to the Bowery Hotel. She paid $6.50 for the ride. She did not tip*. In the future, small data elements gathered in a narrative like this will be generated by a constellation of devices carried by the user, and by her supporting services. A data-rich payment ecosystem based on NFC (near field communication)-enabled devices will create a record of the payment and harvest ride details automatically. These data will then feed into personal applications such as automated diaries, personal expense reports, and time-management aids.

Though these hypothetical personal applications do not yet exist in the market, this type of small data is generated every minute and has already contributed to documented failures of data protection and control. The taxi ride cited above is real: the actress Olivia Munn traveled from Varick Street to the Bowery Hotel in 2013. In 2014, an enterprising researcher chose to mine public data, and published findings of public interest. The researcher had first noticed that publicly posted photos of celebrities entering and exiting New York City taxis often show legible taxi medallion numbers.[6] Though the government of New York City does make data on individual taxi rides publicly available, it takes care to conceal medallion numbers to protect riders and drivers. Unfortunately, in one large data set, the city implemented this protection ineffectively – through misuse of a hash function – making it possible to associate ride information with specific medallion numbers.

Ridesharing services like Uber are yet another way that these types of data are being generated. Thanks to its use of user-generated location data for pickups, Uber has transformed the use of small data in urban transportation. Like many other shared-service providers, the company is blurring the boundary between customer data, which is used to generate sales, and small data, or personal information, which is used to benefit the user. One could imagine Uber consuming additional user-generated data, such as its users' personal calendars, in order to provide more convenient – and powerful – services. A dark facet of Uber's convenience is the "God view," a (once secret) viewing mode available to Uber employees to track any user. Uber has purportedly used the God view to harass journalists who have written critically about the company.[7] In 2012, Uber infamously published a blog post that tracked what the company called "rides of glory": rides whose timing seemed to indicate passengers had engaged in one-night stands.[8] Given that Uber is generating at least a portion of this personal data, the question arises: should individual users

have the ability to delete personal data stored with such services, or should they learn about how their data are used in order to hold service providers like Uber accountable for abuses?

While these particular privacy violations may not be of great concern to the general public, they illustrate the principle that personal data do not always originate directly with the user. More and more, personal data can be sourced from many different places and can emerge unpredictably. When personal information is turned into small data and made available for individual benefit, it comes burdened with complex provenance; thus, consumers will struggle to control small data they perceive as "theirs." Moreover, as small-data use transforms life-altering, positive realms, such as health care and education, the hazards and conflicting interests involved in data creation could bring additional serious issues to the fore, including data entanglement and data integrity.

There are important limitations to existing designs and models for privacy and control. Several existing approaches to data protection, such as *sanitization*, *cryptography*, and *ownership assignment*, do not address the perspective of small data used by and for the individual. Sanitization is, very broadly speaking, the practice of redacting, aggregating, or adding noise to a collection of data to prepare it for safe release in a privacy-sensitive context. This is the approach that the New York City government took to prevent its taxi-ride data from being used to identify customers; it replaced medallion numbers with cryptographically constructed pseudonyms. As that example shows, data sanitization can be a fragile process. One mistake or unanticipated correlation can lead to unwanted data disclosures.

Another problem with sanitization is the trade-off between privacy and utility. Generally, with an increase in utility comes a decrease in privacy. This tension was strikingly demonstrated by a data set from sixteen MOOCs (massive open online courses) run by MITX and HarvardX on the edX platform.[9] To comply with a federal statute known as the Family Educational Rights and Privacy Act (FERPA), scientists "deidentified" the data set, using a privacy measure called "k-anonymity." Subsequently, these data sets were widely studied by researchers. However, the scientists who produced the data set also discovered that the sanitized data differed in marked ways from the original data set. For instance, in the deidentified data set, the percentage of certified students, or those who successfully completed courses, dropped by nearly one half from the true data set. In this case, protecting privacy could have the drawback of invalidating studies meant to improve instruction quality for students.

In the case of small data, the privacy-utility trade-off is particularly problematic, though not unique to it. There are many big-data analyses, such as medical studies, that can be done more or less safely using sanitized data.[10] Sanitization, however, often does not scale down to the protection of small data: it is not possible to hide an individual's data within a crowd's when the utility of the data stems from its successful integration with other data pertaining to that individual. This problem is illustrated by a study of personalized medicine in which researchers examined estimates of stable dosages for warfarin, an anticoagulant medication, that were made using patients' genetic markers.[11] Researchers demonstrated that in the standard model for such recommendations, a patient's estimated stable dose of warfarin leaks information about his genetic markers. Sanitizing the dose data – in other words, preventing leakage of genetic information by using standard privacy-protecting tools within the model – does not work.[12] The model consumes a tiny amount of infor-

mation (only two genetic markers), and the information is only sourced from one individual. Further, the cost of strong sanitization could be fatal. Degrading the fidelity of the model could result in inaccurately estimated stable warfarin dosages, which could very likely cause patient deaths.

There is little motivation for sanitization when data are consumed by the individual who produced them, as is sometimes the case for small data. But given how many opportunities now exist for sharing small data, it would be natural to appeal to sanitization as a privacy-preserving tool.

Another technical approach to enforcing data confidentiality is the use of cryptography, particularly encryption. Take the example of medical data, also known as protected health information (PHI), which is a particularly sensitive form of small data. The federal Health Insurance Portability and Accountability Act (HIPAA) promotes encryption of such data. Organizations that properly encrypt data and store keys can, in the case of a breach, claim safe harbor status and bypass breach notifications.

When properly deployed today, encryption is very robust: a standard algorithm, such as the Advanced Encryption Standard (AES), cannot be broken even by a powerful adversary. At first glance, properly implemented encryption seems like a cure-all for confidentiality issues.

But encryption, like sanitization, acts at odds with utility. Encrypted data cannot be computed on. (Theoretical and application-specific approaches to computing on encrypted data exist, but have limited utility in practice.) A system must have access to data in order to process it, and thus, if presented with encrypted data, must be able to decrypt it. Further, if a system has access to the encrypted data, then so, too, does an attacker that breaches the system or steals credentials, such as passwords, from a person with access. While encryption is an alluring technical approach to protecting privacy, it is not a magical, cure-all solution.

Given the limitations of technical measures in the protection of privacy, a call has arisen to appeal to economic protections, and perhaps even stimulate open markets for personal data. This approach, which we here refer to as ownership assignment, is exemplified by computer scientist Alex Pentland's "Reality Mining of Mobile Communications: Toward a New Deal on Data," which urges that users should "own their own data."[13] Old English Common Law encapsulates this idea in three general rights for tangible property: users should control the possession, use, and destruction, or dispersion, of their data. The "New Deal on Data" goes a step further: users should also be able to treat data handlers like banks, withdrawing their data if desired and, as with Swiss banks, storing it anonymously.

This deal, which is grounded in a common sense physical model, is enticing. But data are distinctly different from land or money. Data management is far more complicated, and it defies physical transactional models. An acre or a dollar cannot be arbitrarily replicated by anyone who sees it. Nor can it be mathematically transformed into a new object.

Data, on the other hand, are infinitely malleable. They can arise in unexpected places and be combined and transmogrified in an unimaginable number of ways.

To understand the complexities of data ownership, we might ask: who owns the data you generate when you purchase electronic toys from Amazon or food from FreshDirect? Who owns the information produced by your viewings of movies on Netflix, or videos on YouTube? Who owns the data generated by your Android phone, purchased from Cyanogen, and connected to the T-Mobile network, to say nothing of the "physiological" data generated by third-party software on your Fitbit or Apple Watch?

In a previous issue of *Dædalus* on "Protecting the Internet as a Public Commons," legal scholar Helen Nissenbaum articulated relevant alternatives to property rights through her suggestion that we understand privacy as contextual integrity (the idea that privacy is a function of social norms and the environment in which disclosure occurs). She argues that instead of focusing on ownership assignment, we focus on the right to access.[14] Our essay is an argument for that right, and further, for the embodiment of that right in the data and services markets and architectures that we are investing in as leaders of organizations, designers of products, executors of regulations, and consumers of services.

But even with this formulation of property rights, complications arise. Many small data settings invoke *involuntary hazard*, in which the handling of small data by one person can affect the privacy or rights of another without his or her knowledge or involvement. This can occur either from joint data creation or from interactions between individuals on a given platform. Emails, blog posts, and group photos all implicate people captured or referenced in these media with or without their consent, just as a Facebook "gift" creates a record of the sender and the (potentially unwitting) recipient. Many more forms of involuntary hazard will arise as cameras and sensors proliferate, as small data are increasingly aggregated in the cloud, and as analysis and correlation of small data streams yield new insights. Innocent bystanders in photographs, for example, could also be implicated by data sharing.

Kinship gives rise to a particularly striking example in the small data handling of involuntary hazard. Parenthood – the ultimate act of joint creation – creates shared genetic material among kin. This genetic data provide a long-term window into a person's health prospects and behavioral characteristics. Given the sensitivity of such data, the U.S. Federal Genetic Information Nondiscrimination Act (GINA) of 2008 prohibits the use of genetic information by health insurers and employers. As a result of direct-to-consumer genetic testing, however, some people choose to post their genetic data online in repositories, such as OpenSNP, to help catalyze medical discoveries.[15] This choice impacts their kin and descendants, potentially for many decades. As shown in a study by security and privacy researcher Mathias Humbert and colleagues, genetic data enable strong inferences about the predisposition of people related to carriers of Alzheimer's disease toward developing it themselves.[16]

Of all the problems with privacy and accountability mechanisms described here, the most fundamental challenge is, perhaps, psychological in nature. As with health risks associated with exposure to toxins in the air and water, individuals' welfare in terms of privacy is typically degraded more by cumulative exposure than by acute events. Galvanizing people to address gradual threats is a significant and major challenge, without a simple solution.

Given the challenges we have enumerated, key design decisions made today will determine whether we can foster an equitable future society that, while flooded with small data, respects both the value of individuality and personal rights and preferences. Such a society could empower individuals to improve their well-being, social connections, and productivity through intentional use of their small data, while largely avoiding the harmful side-effects of data sharing, such as loss of privacy and vulnerability to predatory businesses. Amid an explosion in the generation, collection, and analysis of small data, however, as well as a resulting erosion of existing models of rights and control, how can we articulate and navigate the decisions needed to realize this vision?

We believe that it is both critical to take a step back from existing models of small-data use, confidentiality, and control, and to frame and reflect on three foundational questions.

What are the practically realizable roles and rights of the individual in the management of small data? Granting ownership and control to individuals over their small data alone will not enable meaningful stewardship. History has shown that many individuals do not have the time or interest to administer fine-grained policies for data access and use. (Facebook privacy settings continue to baffle the majority of users.)[17] It is increasingly impractical for people even to be aware of what data they have produced and where it is stored. As small data become ubiquitous, confidentiality will become increasingly difficult to protect, and leaks may be inevitable. What practical remedies are there?

We suspect that any workable remedy will foremost recognize that individuals' rights should not end with disclosure, and should instead extend to data use. Thus, policies such as HIPAA, which emphasize confidentiality as a means of restricting data flow, will need to be supplemented by rights protections that encompass disclosed data and create fair use and accountability. Consider, again, the example of GINA: if people publish their genetic data, their kin should remain protected.

What fundamental bounds and possibilities exist in data privacy and accountability for small data? There will always be trade-offs between utility and confidentiality. As described above, encryption and sanitization can achieve confidentiality, but often at the expense of the data's usefulness. While emerging cryptographic technologies (such as secure multiparty computation and fully homomorphic encryption) have basic limitations and probably will not alter the landscape for many years to come, they delineate possibilities.[18] They show, for example, that it is possible to mathematically simulate a "trusted third party" that discloses only preagreed-upon results of computation over data without ever revealing the underlying data. Trusted hardware such as the pending Intel SGX technology holds similar potential, but with much more practical, medium-term promise.[19]

Access in such a trusted third-party model can be time-bounded – granted for past data, present data, and/or future data – and limited according to any desired criterion or algorithm. Such a permissions-based model is applicable to both streaming and static data, and is especially useful for "activated" linked-data that is long-lived, streaming, and distributed and used in various ways to drive models and algorithms. This model can create a high degree of accountability by constraining and recording the flow of data.

A simulated trusted third party can offer richer options than releasing sanitized data to researchers. For example, the MOOC data in the HarvardX and MITx study could be made available to researchers not as a sanitized dataset, but as an interface (an API, or application program interface) to a system that manages the raw data.

Nonetheless, public or semipublic release of data will always exist in society; thus, we ought to understand privacy as contextual integrity, a function of social norms and the environment in which disclosure occurs.[20] This concept points toward a future in which semantics automatically govern the flow of data. An intelligent system could discover and determine when and how to process and release data on behalf of consumers, and when and how to fabricate plausible white lies on their behalf (a key social norm). This would be a boon for data-enriched social engagement that would also help restore control to users.

What market demands, government regulations, industry self-regulation (through standardized terms of service), and social norms will shape

the rights of consumers to have access to their small data? To answer this question, we might look at the striking tensions in the commercial handling of health and fitness data today. Recognizing the growing importance of health-related data, Apple has offered HealthKit, an app that serves as a hub for such data drawn from mobile apps. At the same time, the company is treating personal data like a hot potato: Apple does not store or otherwise access its users' health data, leaving this task and liability with app developers. Meanwhile, app developers are ravenously collecting "fitness" data that are likely to function as health data in the future. For example, researchers have shown strong correlations between general health and physical movement throughout the day, and many such apps track physical movement. None of these data are being managed under the aegis of HIPAA.[21]

Users often acquiesce to service providers, such as social networks, health and fitness app developers, and online retailers that take possession of their small data and hold it captive, not sharing it with the users themselves or facilitating its export. Online superpowers like Facebook maintain control over user data and interactions to the extent of being able to influence voter turnout in national U.S. elections.[22] Will our digital selves be reassembled on our behalf by monolithic service providers? Or will an ensemble of entities instead act in concert, according to individual users' tastes and objectives? For more than a decade, mobile phones were controlled by the mobile service provider; since the emergence of smartphones and app stores, control has shifted to the consumer. Which future should we design for? Should ownership of personal data be an inalienable right, rather than one that can be blithely signed away through a terms-of-service agreement?

As Vint Cerf, coinventor of the Internet architecture and basic protocols, has remarked, privacy as we conceive of it today is a historical anomaly, possibly born of the urban revolution.[23] In an age of selfies and social networks, there is every reason to believe that the individual's notions of boundaries and privacy, and of what constitutes personal and public, will continue to shift. Explicit models of the social norms driving policy and practice, and their relationships with market forces and government regulation, must be central to the project of designing the architecture of next-generation small-data systems.

Building methods, tools, and systems with small data in mind is an explicit design choice. By recognizing the role of the individual as beneficiary of her own data-driven applications and services, we are choosing to consider design criteria that are different from those faced by service providers.

If we build systems and market practices that routinely provide people with direct programmatic access to their small data, along with the ability to export and use it, applications and services can offer users the benefit of highly individualized modeling and recommendations that would neither be possible nor acceptable otherwise. And yet, in building such systems, how do we also provide consumers with the safeguards to manage small-data exposure and its consequences in the long term, while still maximizing individual benefit?

We have presented what we believe to be the core challenges raised by small data. We hope that posing these questions is a first step in the direction of secure and beneficial use of small data – by individuals, governments, and enterprises alike.

ENDNOTES

* Contributor Biographies: DEBORAH ESTRIN, a Fellow of the American Academy since 2007, is Professor of Computer Science at Cornell Tech and Professor of Healthcare Policy and Research at Weill Cornell Medical College. She is Founder of the Jacobs Institute Health Tech Hub and Cofounder of the nonprofit Open mHealth. Her recent publications include articles in *Journal of Medical Internet Research*, *Journal of Acquired Immune Deficiency Syndromes*, and *ACM Transactions on Intelligent Systems and Technology*.

ARI JUELS is Professor at the Jacobs-Technion Cornell Institute at Cornell Tech. He has recently published articles in *Journal of Cryptology*, *Communications of the ACM*, and *IEEE Security & Privacy Magazine*.

1 Deborah Estrin, "Small Data, Where *n* = Me," *Communications of the ACM* 47 (4) (2014): 32 – 34.

2 See the collaboration between users and manufacturers of self-tracking tools at http://quantifiedself.com/.

3 See https://www.patientslikeme.com/; and https://corp.inspire.com/patients-caregivers/.

4 For an example of such terms of use, see MyFitnessPal, "Terms of Use," http://www.myfitnesspal.com/account/terms_and_privacy?with_layout=true (accessed January 23, 2015).

5 "Fighting for the Right to Open His Heart Data: Hugo Campos at TEDxCambridge 2011," TEDx Talks, uploaded January 19, 2012, https://www.youtube.com/watch?v=oro19-l5M8k.

6 J. K. Trotter, "Public NYC Taxicab Database Lets You See How Celebrities Tip," *Gawker*, October 23, 2014, http://gawker.com/the-public-nyc-taxicab-database-that-accidentally-track-1646724546.

7 In fact, Emil Michael, Uber's senior vice president of business, had suggested to a private audience that the company might dedicate a portion of its financial resources to private researchers to investigate adversarial journalists' personal lives in retaliation for their negative coverage. See Gail Sullivan, "Uber Exec Proposed Publishing Journalists' Personal Secrets to Fight Bad Press," *The Washington Post*, November 18, 2014, http://www.washingtonpost.com/news/morning-mix/wp/2014/11/18/uber-exec-proposed-publishing-journalists-personal-secrets-to-fight-bad-press/; and Chanelle Bessette, "Does Uber Even Deserve Our Trust?" *Forbes*, November 25, 2014, http://www.forbes.com/sites/chanellebessette/2014/11/25/does-uber-even-deserve-our-trust/.

8 Bessette, "Does Uber Even Deserve Our Trust?"

9 For example, Munn and other celebrities whose rides surfaced in this data-mining exercise were criticized for not tipping their taxi drivers. Some alleged, though, that the taxi drivers themselves intentionally failed to record tips. In other words, Ms. Munn's small data may have been corrupted by a "privacy-conscious" (a euphemism for "tax-evading") taxi driver.

10 Jon P. Daries, Justin Reich, Jim Waldo, Elise M. Young, Jonathan Whittinghill, Daniel Thomas Seaton, Andrew Dean Ho, and Isaac Chuang, "Privacy, Anonymity, and Big Data in the Social Sciences," *ACM Queue* 12 (7) (2014), http://queue.acm.org/detail.cfm?id=2661641.

11 Benjamin C.M. Fung, Ke Wang, Rui Chen, and Philip S. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments," *ACM Computing Surveys* (CSUR) 42 (4) (2010), doi:10.1145/1749603.1749605.

12 Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart, "Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing," *Proceedings of the 23rd USENIX Security Symposium* (Berkeley: USENIX, 2014), https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/fredrikson_matthew.

13 Cynthia Dwork, "Differential Privacy: A Survey of Results," in *Theory and Applications of Models of Computation*, ed. Manindra Agrawal, Dingzhu Du, Zhenhua Duan, and Angsheng Li (Berlin: Springer Berlin Heidelberg, 2008), 1 – 19.

14 Alex Pentland, "Reality Mining of Mobile Communications: Toward a New Deal on Data," in *The Global Technology Report 2008 – 2009: Mobility in a Networked World*, ed. Soumitra Dutta and Irene Mia (Geneva: World Economic Forum, 2009).

15 Helen Nissenbaum, "A Contextual Approach to Privacy Online," *Dædalus* 140 (4) (Fall 2011): 32 – 48.

16 OpenSNP, https://opensnp.org.

17 Mathias Humbert, Erman Ayday, Jean-Pierre Hubaux, and Amalio Telenti, "Addressing the Concerns of the Lacks Family: Quantification of Kin Genomic Privacy," *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security* (2013): 1141 – 1152, doi:10.1145/2508859.2516707.

18 Yabing Liu, Krishna P. Gummadi, Balachander Krishnamurthy, and Alan Mislove, "Analyzing Facebook Privacy Settings: User Expectations vs. Reality," *Proceedings of the ACM SIGCOMM Conference on Internet Measurement* (2011): 61 – 70, doi:10.1145/2068816.2068823.

19 Craig Gentry, *A Fully Homomorphic Encryption Scheme*, Ph.D. dissertation for Stanford University Department of Computer Science (September 2009), https://crypto.stanford.edu/craig/craig-thesis.pdf; and Marten van Dijk and Ari Juels, "On the Impossibility of Cryptography Alone for Privacy-Preserving Cloud Computing," *HotSec 2010 Proceedings of the 5th USENIX Conference on Hot Topics in Security* (Berkeley: USENIX, 2010), 1 – 8.

20 Ittai Anati, Shay Gueron, Simon P. Johnson, and Vincent R. Scarlata, "Innovative Technology for CPU Based Attestation and Sealing," *Proceedings of the 2nd International Workshop on Hardware and Architectural Support for Security and Privacy* (June 2013).

21 Nissenbaum, "A Contextual Approach to Privacy Online."

22 Robert Ross and K. Ashlee McGuire, "Incidental Physical Activity is Positively Associated with Cardiorespiratory Fitness," *Medicine and Science in Sports and Exercise* 43 (11) (2011): 2189 – 2194.

23 Zoe Corbyn, "Facebook Experiment Boosts U.S. Voter Turnout," *Nature News*, September 12, 2012, http://www.nature.com/news/facebook-experiment-boosts-us-voter-turnout-1.11401.

24 Gregory Ferenstein, "Google's Cerf Says 'Privacy May Be An Anomaly.' Historically, He's Right," TechCrunch, November 20, 2013, http://techcrunch.com/2013/11/20/googles-cerf-says-privacy-may-be-an-anomaly-historically-hes-right/

Choices: Privacy & Surveillance in a Once & Future Internet

Susan Landau

Abstract: The Internet's original design provided a modicum of privacy for users; it was not always possible to determine where a device was or who was using it. But a combination of changes, including "free" Internet services, increasing use of mobile devices to access the network, and the coming "Internet of Things" (sensors everywhere) make surveillance much easier to achieve and privacy more difficult to protect. Yet there are also technologies that enable communications privacy, including address anonymizers and encryption. Use of such technologies complicate law-enforcement and national-security communications surveillance, but do not completely block it. Privacy versus surveillance in Internet communications can be viewed as a complex set of economic tradeoffs – for example, obtaining free services in exchange for a loss of privacy; and protecting communications in exchange for a more expensive, and thus less frequently used, set of government investigative techniques – and choices abound.

SUSAN LANDAU is Professor of Cybersecurity Policy in the Department of Social Science and Policy Studies at Worcester Polytechnic Institute. Previously, she served as Senior Staff Privacy Analyst for Google and as Distinguished Engineer at Sun Microsystems. She is the author of *Surveillance or Security? The Risks Posed by New Wiretapping Technologies* (2011) and *Privacy on the Line: The Politics of Wiretapping and Encryption* (with Whitfield Diffie, 1998; rev. ed. 2007), and contributed to the National Research Council's *Bulk Collection of Signals Intelligence: Technical Options* (2015).

Electronic communications create challenges. In enabling citizens to connect at a distance, they would appear to loosen governmental control. But signals can be eavesdropped on and recorded, and communications surveillance gives tremendous power. Even if the communication itself is encrypted, communications metadata – the who, when, where of a message – are not. Anyone who can collect metadata has vast opportunity to know who is where and connecting with whom.

The ability to remotely eavesdrop has existed for at least as long as electronic communications. Because a signal can be plucked from the air, rather than visibly tapped into, radio is easier to eavesdrop on than are wired communications. In many ways, the Internet has made such surveillance easier still. The tremendous flexibility afforded by the network – the medium supports applications as diverse as search engines, maps, online social networks (OSNs), Twitter, YouTube, Netflix, Uber, and MOOCs (Massive Open Online Courses) – makes the Internet indispensable to citizens and nations alike, and its signals

doi:10.1162/DAED_a_00365

provide rich content for anyone listening in. This can include governments, suppliers of the services, and eavesdroppers.

Yet it does not have to be that way. Communications can be encrypted, and, like speech, they can be ephemeral. The record of a communication's path through a network can be essentially undiscoverable. There are many ways to provide Internet communications, some of which do not impinge on privacy. This article is about those choices.

It starts, as it often does, with the underlying technology. The Internet was developed as a medium for sharing data. Its basic architectural principles – to break data into a numbered set of small packets and transmit the packets as efficiently as possible – reflects that underlying premise. Each packet is transmitted using the Internet protocol (IP). Packets typically have three parts. The *header* says where the packet is from (the sender's IP address), where the packet is going (the receiver's IP address), the type of communications protocol (email, Web page, video, voice, and so on), and its position (packet number) in that particular transmission. The *payload* – the actual content – follows. Finally a *trailer* marks the end of a packet. Applications – an http connection to a Web page, an email connection, a Voice over IP call (VoIP) – are broken into packets and then reassembled at the receiver's end.

Mobility of devices means that the user's IP address at the café at 10 a.m. is different from that in the seminar room at 11 a.m. Each time the user connects back to the network, her IP address is transmitted to her service provider. That is how Facebook communications and your email reach her even when she has moved locations and her IP address has changed.

IP location provides partial identification. While an IP address delimits a location from which and to which packets are transmitted, that address is, for a number of reasons, not necessarily useful in identification. The IP address may be one used temporarily, and without strong identification, such as at an Internet café or an airport. Without ancillary information, such an IP address may provide minimal identifying information. Another reason that an IP address may not provide definitive identification is that few routers along the transmission check a sender's address; so spoofing an IP address is easy.[1]

Even if the IP address is correct, it may not provide an investigator with information to determine who is responsible for a particular action.[2] That is because in such instances, the connecting machine may be just a way station. Consider, for example, *DDoS attacks* (Distributed Denial of Service attacks), in which hundreds of thousands of computers simultaneously send messages to an online service, overwhelming it and taking it offline. The machines sending these messages are simply intermediaries that have been compromised themselves. DDoS is an example of a *multistage attack*, in which a perpetrator infiltrates a series of machines to launch an attack. *Cyberexploits* – theft of information from networked systems – are also typically multistage attacks. The first machine to receive the exfiltrated data is often itself compromised, and the stolen data will be quickly moved from that machine to another and another – a lengthy chain of compromised machines – before the data end up in the attacker's hands. An investigation may lead to the initial machine that was used in the scam, but is unlikely to lead all the way to the real attacker.[3]

The fact that IP addresses do not provide precise identity matters very little in certain cases. Spoofing does not concern the Recording Industry Association of America, which uses an IP address as a jumping-off point for copyright infringement suits.[4]

IP addresses have also served law enforcement as a starting point for investigations.[5] They can also be useful in investigations in which the participants' addresses are related; for example, if they all work at Enron.

Since an IP address is typically not enough to identify an individual, a user browsing generic sites such as *The New York Times*' without logging in achieves some anonymity. Actions the user takes, however, can alter that. In particular, a series of sufficiently personal searches that can be linked to each other may suffice to identify an individual.[6]

The point is that IP addresses have a fungibility, at least when it comes to identity. They provide a starting point for linking a person with an action, but they are also insufficient to be definitive. Thus, IP addresses can provide surveillance capabilities and privacy; the specific circumstances determine which it might be.

Investigators often seek identity, though not necessarily at the level of an IP address. Following users across the Internet became important with the arrival of free services such as Facebook, Google search, and Yik-Yak. Such services are supported through advertising. In this instance, identity does not mean identifying a user in the sense of "Alison Clark is visiting honda.com," but rather that the browser currently viewing nytimes.com is the same that ran an Internet search for compact cars earlier in the day. This enables the search provider, for example, to serve a Honda ad on *The New York Times* website that the user is browsing. Identification is derived through cookies in the browser, not an IP address.

There are times when identity on the Internet at the level of a person matters. A bank does not particularly care what a user's IP address is, but if there's a transaction occurring, the bank seeks assurance that the person is who she claims to be and wants her to authenticate her identity at the bank's site. For many situations, including transactions with high value, authentication conducted within an application is sufficient proof of identity.

Increasingly, identity is required for accessing services. *Federated identity management* – facilitating access to different sites once a user has been authenticated to an "identity provider" – is one way to do this. For example, a corporate login could permit seamless electronic access to outsourced services such as HR or travel booking; a university login could allow access to electronic resources at a federated institution.

Some approaches to identity management carefully protect privacy. One example is Shibboleth, which is used for sharing secured Web resources and services among a consortium of universities, research labs, publishers, and medical libraries. To access a resource, the user must establish her right to it, such as by being a faculty member or a student. The user's ID is shared only if access to the resource requires it.[7] Another case of privacy-protective identity management comes courtesy of the U.S. government, which employs private-sector identity providers for accessing government websites, but requires that the users' information be employed only for authentication, audit, and complying with the law – and not for ads or sharing with third parties.[8] So if a user is looking at Veterans Affairs benefits and then at information about sexually transmitted diseases, that information should neither be tracked nor stored by the identity provider.

Other systems take a very different approach, using user data to entice services to work with them. Thus, for example, when the Facebook login is used to authenticate a user to an app, Facebook shares with the app the user's name and gender, and provides a list of the user's friends who also use the application. This makes the Facebook login valuable to the app, but not to a user seeking privacy.

The existing model of advertising and tracking in exchange for services is not the only possible model for the Internet. One alternative would be to charge for services: a tenth of a cent for a search, a monthly charge for email support, and so on. And there is no reason the two systems could not coexist: charges for users seeking privacy-protective services, and an advertise-and-track model for those who are indifferent to the privacy issue or unable to pay.

By making the network indispensable to daily life, the Internet drove the development of smartphones. Most Internet accesses now occur through mobile devices, a fact with profound implications for privacy and surveillance. While a laptop can be "on" but not connected to the network – functioning as a computer, not a communications device – if a smartphone is on (and not in "airplane mode"), it will be connected to the telephone network whenever the provider's system is within range. Thus, a phone's location, which is broadcast several times an hour to announce "I am here," is a relatively public piece of information. The phone's connection is through the nearest base station: the cell tower closest to the user. As the user moves to new locations, the phone connection is "handed off" to the next base station. That is information that the phone network – or an interceptor – will learn.

Where an individual is calling from, or whom they called, may be much more interesting than what they actually said; communications metadata, for example, can reveal the structure of an organization. One striking example of this type of analysis concerns the case of former Lebanese Prime Minister Rafik Hariri, who was assassinated in Beirut in 2005, when a truck bomb exploded near his motorcade.[9] The planning behind the assassination was well-hidden, but analysis of cell phone traffic in Beirut and other locations exposed a pattern of communications that revealed who did it – and how.[10]

Desktops, laptops, and tablets are, to some extent, multiuser machines; but smartphones are more strictly associated with individuals. Thus, just tracking the phone's location provides an extremely accurate way of determining a phone's user.[11] Know the recipients of a person's calls, and you can infer who she is and what is happening in her life: whether she has just lost her job, her mother is ill, or her son has just gone off to college. Because people carry personal transmitters and receivers, government investigators no longer need to tail individuals and monitor phone booths to capture conversations and movements; they simply track mobile phones. Because communications patterns are so revealing, if a government can fully surveil a nation's communications network, it can even track "burner" phones (anonymous prepaid phones) through correlations in location and use.

Governments are not the only ones following users' locations; in fact they may be collecting far less information than many companies. To provide the Internet with services for which smartphones are valued – finding a local restaurant and making dinner reservations and then determining the best route there – the phone must provide location information to the app. This is done through GPS, which typically operates on a resolution within ten meters.[12] So the network provider knows where the phone is and with which service the user is communicating, while the app provider learns phone location and what information is delivered through the app.[13]

This is an interesting design choice in location data tracking: Apple's iOS8 does not allow apps to collect location information when the app is not in use, but there is no such restriction for Android phones (of course, if location collection is shut off, then Android apps cannot collect it). The

latter situation might change: in February 2015, the U.S. Federal Trade Commission told app developers, "If you access users' locations when they're not using your app, it's a good idea to clearly disclose what you're doing and provide users with choices."[14]

The real gold in the Internet advertising world is "conversion tracking," learning what customers do after clicking on an ad: whether they bought the product or followed up in another way (such as visiting a product's website). When the Web access, user location, and payment are all on the same device, it becomes even easier for an Internet service and an advertiser to determine an ad's effectiveness. The phone might not announce, "This is Alison Clark" at the Honda dealership, but if her phone shows an identifier from the search she conducted, that provides the relevant information. For this reason, companies are at least as eager as governments to use smartphones to track users.

With such interest in following the user and such capabilities for surveillance, it becomes difficult to imagine that any privacy is possible. Yet there are many technical solutions for protecting privacy. It is particularly striking that there are even technical solutions for obscuring with whom you are communicating. In the mid-1990s, the Naval Research Laboratory began work on a system that makes it difficult to determine who is connecting with whom on the Internet.[15]

The onion routing network, commonly known as Tor, protects against traffic analysis through deployment of a "Tor network," a collection of servers with encryption and decryption software. A path is determined for each communication, which is then routed through a network of Tor nodes (servers) that strip off the encryption "one layer at a time." Encryption keys are based on the nodes and route.

Anyone who is eavesdropping on Internet traffic can determine that one Tor node is communicating with another. More specifically, if there is surveillance of connections to a website – such as who in Iran is reading about international sanctions – the interceptor will see a connection from the Tor network to the forbidden website. But the eavesdropper will not see the IP address that initiated the Internet connection unless the eavesdropper can view the entire network at once and thus correlate times and sizes for all network transmissions. In such cases, interceptors can deanonymize Tor communications, but otherwise Tor makes such identification extremely difficult.[16] Browsers and instant messaging apps can be used on the Tor network, enabling truly anonymous communication through which it is infeasible for the receiver to determine the original sender's IP address.

It might be surprising that a U.S. government agency supports anonymous Internet accesses. But there are good reasons for the government to seek such capabilities. A military unit in a safe house in the Middle East would not want to let the local Internet service provider (ISP) know that it is communicating with the Naval Academy in Annapolis, while an FBI agent investigating a child pornography chat room does not want to use an IP address that resolves to "fbi.com." So a system that makes it appear that the Web connection is from somewhere else provides useful investigative capability. Tor is widely available and popular with journalists, human-rights workers, and others seeking privacy of communication. And it provides cover for military personnel and law-enforcement investigators, whose identities as U.S. government employees are masked by the system's broader set of users.

In many ways, confidentiality of communications is simpler to achieve than privacy. Encryption – encoding messages so that only the sender and receiver can read

it – accomplishes this. But simple answers belie simple understanding. By now it should be clear that nothing about protecting communications is entirely simple.

For a quarter-century, from the 1970s to the late 1990s, the U.S. government battled academia and industry over encryption used to support confidentiality. This fight came to a head during the "Crypto Wars" two decades ago, at the dawn of the Internet era. In 1999, the European Union loosened its controls on the export of products with strong encryption; a few months later the United States did the same.[17] This change made it much simpler to deploy cryptography in commercial products.

While use of encryption for confidentiality had been controlled, its use for authentication – assurance that a person or site is who they say they are – had not. Https, the secure version of the http linking protocol, is used to authenticate a website (for example, confirming that the site is bankofamerica.com and not an imposter that is like bankofamerica.com) and encrypt communications between a user and the site. This protocol was essential for electronic commerce, and was already deployed by the mid- to late 1990s. Given that https was widely deployed quite early for ecommerce, it is surprising and somewhat striking that Web mail, the service that provides email through a browser, was not similarly protected. Let us examine how such services work.

Suppose a user with the email account boris@yahoo.com is communicating with another user, natasha@gmail.com. When Boris sends an email to Natasha, he logs on to his Yahoo! Mail account, writes his message, hits send, and the mail travels to Natasha's Gmail account. She will read his message once she logs onto the Gmail server (many users, including those with Android phones, are always logged on).

From the beginning, Web mail providers used the https protocol when authenticating users to their accounts; this encrypted the user password from the user's browser to the site. But for many years, the large Web mail providers – Gmail, Hotmail, Yahoo! Mail – did not encrypt the connection between the user and her mail account; that is, *the emails themselves traveled in the clear* between the user's machine and the provider. Anyone eavesdropping on the Internet connection between Boris and his Yahoo! Mail account, relatively easy to do, could read Boris's incoming and outgoing mail. In response to the Snowden revelations, Google changed their connection to a secure one, and other providers are following suit. Mail traveling between Natasha's computer and her Gmail account are on an encrypted channel; interceptors cannot read it.

But this change does not fully encrypt the mail from Boris to Natasha. Although systems began securing communications between the user and the mail provider, the communications themselves still are not encrypted "end-to-end" from sender to receiver. If Boris and Natasha both happen to be using Gmail, then their communication will be encrypted between Boris and the Gmail server and between the Gmail server and Natasha. Contents on the Gmail servers are encrypted, but there will be a time when Boris's mail to Natasha is in the clear at Google. That is because Google uses the mail to serve ads and to provide personalized services. For example, a plane reservation in an email account will trigger a notification in the Google Now app to inform the user about traffic on her preferred route to the airport.

There are other models for email, some of which provide greater confidentiality. One such service was Lavabit. Mail on Lavabit servers arrived encrypted and stayed that way; they were decrypted only when a user was reading the communication. Users received keys through a secure https connection.

Lavabit was shut down by owner Ladar Levison after the U.S. government requested the encryption key securing the https connection between Lavabit and its users. Although government investigators appeared to be interested only in a single user's communications, giving up that key would have allowed access to all https connections, thus potentially exposing all customer passwords. Levison felt that would violate his privacy commitment to his customers. Instead of doing so, he closed the service.[18]

Another example of alternative privacy protection is Off-the-Record (OTR) chat. Google's OTR chat does not store chat histories in users' accounts, or in the accounts of the people with whom they are chatting. But Google policy does not preclude storing the communications elsewhere.[19] A more protective version would be not to store the communications whatsoever. Even more protective would be not storing and providing encryption for the chat. Most protective would be to encrypt using a technique called *forward secrecy*, so that even if the encryption key is compromised at some point, no previously intercepted messages can be decrypted.[20] There are OTR systems that provide this level of security.[21]

Alternatives in designing applications lead to varying degrees of privacy. Such safeguards do not come for free. They cost extra development time and can decrease efficiency by preventing reuse of data in other applications. And, as Lavabit's owner discovered, sometimes privacy-protective systems lead to conflicts with the government.[22]

Encryption's knotty issue is that legal access to decrypted content may be granted to an investigator, but technology prevents such access. And although electronic communications now provide much richer investigative information than ever before – consider the Hariri case – sometimes content provides information that these other tools cannot. There is, however, a way around this problem.

As the Snowden disclosures confirmed, national-security agencies may exploit vulnerabilities in communications devices to exfiltrate data from targets.[23] Such capabilities are used not only by intelligence agencies, but by law enforcement as well.[24] As encryption becomes increasingly common, such "lawful hacking" will increasingly be used when communications content cannot be retrieved in other ways.[25] It is no silver bullet; a vulnerabilities approach is more complex legally and technically, and more expensive than if unencrypted communications can be made available.

The privacy situation is about to grow far more difficult. While Internet transactional information is remarkably revelatory, the information from sensors on toothbrushes, watches, clothes, heart monitors, phones – and everything else – will be many times more so. Cheap sensors communicating with the Internet will soon be everywhere: sensors to measure tire pressure and bridge structural health; sensors to report on the freshness of food in the fridge, the dampness in the soil, and the movement of an elderly person at night; sensors to determine whether the car driver or passenger is making a call. The number of devices from the Internet of Things (IoT) will dwarf the current number of devices connected to the Internet.

A user has some control over whether information on her smartphone is shared with the app; she can always shut the application off or completely remove it from her phone. While in some cases – as with smart toothbrushes – the user might have the same capability, she is unlikely to be provided with such control on many other applications (such as tire sensors).

For security's sake, one approach might be enclaves: creating domains with extreme-

ly limited ability to communicate outside a narrow realm. Consider the type of connectivity a smart refrigerator should have. Fully connecting to the Internet creates an unnecessary security risk. The fridge needs to be able to communicate with the manufacturer for updates and with the owner for the you-need-milk notifications. A smartphone app that puts milk on the shopping list does not need notifications directly from the fridge; it could do so instead by accessing owner updates. Limiting information flows from sensors and controlling where those data initially go provides a measure of privacy and security.

Enclaves are likely to be for systems of similar purposes (medical devices, infrastructure sensors). A patient-sensor network in a hospital intensive care unit should not be accessible outside that area, while a sensor network for medical research might span wide geographies. In some cases, data can be aggregated before reaching a larger network – such as combining data from sensors on soil conditions within a region – providing privacy to individuals. Flows of information – which data are shared with whom – will be determined by enclaves.

Determining appropriate enclaves for sensor networks – should the enclave for medical research networks be strictly separate from that used for patient networks? – is complex, but provides only a partial solution for privacy. This is partly because keeping enclaves truly separate remains a difficult technical problem. "Car-hack" attacks, such as when in 2010 researchers remotely took control of a car's brakes and engine,[26] were possible because enclaves lacked clear separation. In addition, putting tight legal and policy controls on the data's use will be crucial for privacy.[27]

Communications between people at a distance have never been entirely private. Delivery is variable, seals can be broken, messages decrypted.[28] Communications that were once ephemeral now have a trail, and being anonymous in modern society is no longer plausible. It not only means eschewing the use of smartphones (and credit cards, transit passes, and so forth), but also requiring companions to do the same. You cannot hide from network detection if your known companions' phones broadcast their whereabouts.

In the wake of the Snowden disclosures, privacy-enhancing technologies such as Tor, and Google and Apple's encrypted phones, in which decryption is only possible with the user key (though, of course, much of the data may also be stored elsewhere), have drawn much interest. Privacy-enhancing technologies enable different levels of ability to conceal identity and increase the cost for monitors to determine information about an individual, but data collection is so vast that these tools are unlikely to be sufficient for people with specific needs to protect themselves, including journalists and human-rights workers, as well as criminals, terrorists, and spies. Indeed, serious efforts to defend against electronic traces may only draw increased attention from intelligence agencies or other eavesdroppers.[29]

Privacy has always been about economics. How much does it cost to use Lavabit's encrypted email services instead of free Gmail services? Or how much more does it cost to use cash at the bookstore instead of ordering over Amazon? On the flip side, how many resources must be devoted to investigations if communications are protected through privacy-protective technologies?

The Internet changed the equation in various ways. In the initial development of Internet applications, we tipped in one direction, allowing collection and release of massive amounts of information about ourselves. Application design, however, provides a plethora of possibilities. As

long as "free" is the model for Internet services, it is unlikely that the tracking industry, developed to support Internet advertising, will disappear. The information amassed by private industry, including the vast collection of data afforded through the Internet of Things, will also be accessed by governments.

Our current Internet design is a world in which applications sometimes provide privacy-protective solutions for those who want them. But these give only a modicum of privacy. Changing the ease with which surveillance can be performed, making it more difficult to track user preferences and activity, is largely a matter of choices. (Of course, under some governments, there are no such choices. But in the United States, private industry is not required to know who users are in order to provide them a service.) Choices for more privacy-protective solutions can come from government regulations, and they can come from customer demand. But such alternatives in application design do exist.

Humans are a highly communicative species, and the Internet fed this aspect of our nature. That the Internet grew spectacularly alongside the terrorist attacks of September 11th and their aftermath meant that privacy, always on a societal pendulum, largely suffered over the last decade and a half. Now choices abound; we may be reaching a time when the pendulum swings back. But the market will only provide effective privacy-protective solutions if enough users demand them.

ENDNOTES

1 Robert Beverly, Ryan Coga, and kc claffy, "Initial Longitudinal Analysis of IP Source Spoofing Capability on the Internet," July 25, 2013, http://calhoun.nps.edu/bitstream/handle/10945/36775/Beverly_Initial_Longitudinal_2013.pdf.

2 This discussion on the value of IP addresses for attribution is based on David Clark and Susan Landau, "Untangling Attribution," *Harvard National Security Journal* 2 (2) (2011): 25 – 40.

3 In the case of DDoS attacks, the machine sending connection requests to the service has itself been infected with malware. By examining ISP logs, it will usually be possible to determine from where it is receiving instructions to attack (flood the service with connection requests). But determining which machine, or set of machines, are sending these instructions may be a multistep process, since instructions may be laundered through several machines. A similar situation exists for data exfiltration, with the wrinkle that the data can be followed only until they reach a jurisdiction in which the trail is made opaque. For further discussion, see ibid.

4 This technique was successfully used by the RIAA for a number of years, although determining an infringing user from an IP address is not straightforward; see *Recording Industry Association of America, Inc. v. Verizon Internet Service*, 351 F. 2nd 1229, D.C. Circuit (2003). Since 2012, a number of rulings have gone against the RIAA.

5 See Clark and Landau, "Untangling Attribution," fn 3.

6 AOL released information about users' searches over a three-month period; identifying a particular user was not hard to do. See Michael Barbaro and Tom Zeller Jr., "A Face is Exposed for AOL Searcher No. 4417749," *The New York Times*, August 9, 2006. The ability to link a set of searches to a user requires first being able to link the user to her searches.

7 R. L. Morgan, Scott Cantor, Steven T. Carmody, Walter Hoehn, and Kenneth J. Klingenstein, "Federated Security: The Shibboleth Approach," *EDUCAUSE Quarterly* 27 (4) (2004): 12 – 17.

8 Georgia Tech Research Institute, "GTRI NSTIC Trustmark Pilot" (October 7, 2014), https://trustmark.gtri.gatech.edu/operational-pilot/trustmark-definitions/ficam-privacy-activity-tracking-requirements-for-csps-and-bae-responders/1.0/.

9 The case is currently being decided in the Hague by the UN's Special Tribunal for Lebanon.

10 The cell phone traffic showed several groups coordinating actions while tracking Hariri through Beirut, and possibly even conducting a dry run of the attack. See Ronen Bergman, "The Hezbollah Connection," *New York Times Magazine*, February 10, 2015.

11 Phillippe Golle and Kurt Partridge, "On the Anonymity of Home/Work Location Pairs," in *Pervasive Computing : 7th International Conference, Pervasive 2009, Nara, Japan, May 11 – 14, 2009, Proceedings*, ed. Hideyuki Tokuda, Michael Beigl, Adrian Friday, A. J. Bernheim Brush, and Yoshito Tobe (Berlin : Springer Berlin Hedelberg, 2009), 390 – 397.

12 Matt Blaze, Testimony to the House Committee on the Judiciary Subcommittee on the Constitution, Civil Rights, and Civil Liberties, *Hearing on ECPA Reform and the Revolution in Location Based Technologies and Services*, June 24, 2010.

13 If the user is signed in, as are most Android users, then Google will learn which apps are being used and how frequently, though not what information is being communicated (unless the apps are Google apps).

14 See Amanda Koulousias, "Location, Location, Location," Federal Trade Commission, February 11, 2015, https://www.ftc.gov/news-events/blogs/business-blog/2015/02/location-location-location.

15 See https://www.torproject.org.

16 James Ball, Glenn Greenwald, and Bruce Schneier, "NSA and GCHQ Target Tor Network that Protects Anonymity of Web Users," *The Guardian*, October 4, 2014.

17 Export controls had effectively prevented the deployment of cryptography in domestic products. While the change in regulations did not permit export of cryptography in all products, it worked well enough to support the needs of the expanding Internet ecosystem. See Whitfield Diffie and Susan Landau, "The Export of Cryptography in the 20th Century and the 21st," in *The History of Information Security : A Comprehensive Handbook*, ed. Karl De Leeuw and Jan Bergstra (Amsterdam : Elsevier, 2007), 725 – 736.

18 See Nicole Perlroth and Scott Shane, "As F.B.I. Pursued Snowden, an E-Mail Service Stood Firm," *The New York Times*, October 2, 2013.

19 Google Support, Chat Help, "Chatting Off the Record," https://support.google.com/chat/answer/29291?hl=en (accessed March 29, 2015).

20 Whitfield Diffie, Paul van Oorschot, and Michael Wiener, "Authentication and Authenticated Key Exchanges," *Designs, Codes, and Cryptography* 2 (2) (June 1992) : 107 – 125.

21 Surveillance Self-Defense, "How to : Use OTR for Mac," and "How to : Use OTR for Windows," Electronic Frontier Foundation, https://ssd.eff.org/en/index (accessed March 30, 2015).

22 Levison had previously complied with court orders for targeted access. His objection to the FBI request was that the agency sought the encryption key for his SSL certificate, which would have compromised the privacy of all Lavabit users. See Perlroth and Shane, "As F.B.I. Pursued Snowden, an E-Mail Service Stood Firm."

23 Spiegel Staff, "Inside TAO : Documents Reveal NSA Top Hacking Unit," *Spiegel Online International*, December 29, 2013.

24 In this case, to determine IP addresses ; but the method can also be used to exfiltrate data, including encryption keys. See Steve Bellovin, Matt Blaze, Sandy Clark, and Susan Landau, "Going Bright : Wiretapping without Weakening Communications Infrastructure," *IEEE Security and Privacy* 11 (1) (January/February 2013) : 62 – 72 ; and Steve Bellovin, Matt Blaze, Sandy Clark, and Susan Landau, "Lawful Hacking : Using Existing Vulnerabilities for Wiretapping on the Internet," *Northwestern Journal of Technology and Intellectual Property* 12 (1) (2014).

25 See Bellovin et al., "Going Bright."

26 Karl Koscher, Alexei Czeskis, Franziska Roesner, et al., "Experimental Security Analysis of a Modern Automobile," in *Proceedings of IEEE Symposium on Security and Privacy (Oakland) 2010*,

ed. David Evans and Giovanni Vigna (Washington, D.C.: IEEE Computer Society, 2010), 447–462.

27 See President's Council of Advisors on Science and Technology, *Big Data and Privacy: A Technological Perspective* (Washington, D.C.: Executive Office of the President, May 2014), https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf.

28 Message decryption led to the downfall of Mary, Queen of Scots, in the sixteenth century; see The National Archives of the United Kingdom, Codes and Ciphers, "Mary's Ciphers," http://www.nationalarchives.gov.uk/spies/ciphers/mary/ma1.htm.

29 The lack of connection to communication networks was one hint that bin Laden was in the villa in Abbottabad; see Mark Mazzeti, Helen Cooper, and Peter Baker, "Behind the Hunt for Bin Laden," *The New York Times*, May 2, 2011.

As the Pirates Become CEOs: The Closing of the Open Internet

Zeynep Tufekci

Abstract: The early Internet witnessed the flourishing of a digitally networked public sphere in which many people, including dissidents who had little to no access to mass media, found a voice as well as a place to connect with one another. As the Internet matures, its initial decentralized form has been increasingly replaced by a small number of ad-financed platforms, such as Facebook and Google, which structure the online experience of billions of people. These platforms often design, control, influence, and "optimize" the user experience according to their own internal values and priorities, sometimes using emergent methods such as algorithmic filtering and computational inference of private traits from computational social science. The shift to a small number of controlling platforms stems from a variety of dynamics, including network effects and the attractions of easier-to-use, closed platforms. This article considers these developments and their consequences for the vitality of the public sphere.

ZEYNEP TUFEKCI is Assistant Professor at the School of Information and Library Science, with an affiliate appointment in the Department of Sociology, at the University of North Carolina, Chapel Hill. She is the author of *Beautiful Tear Gas: The Ecstatic, Fragile Politics of Networked Protests in the 21st Century* (forthcoming 2016) and coeditor of *Inequity in the Technopolis: Race, Class, Gender, and the Digital Divide in Austin* (with Joseph Straubhaar, Jeremiah Spence, and Roberta G. Lentz, 2012). She is also a contributing opinion writer for *The New York Times*.

I traveled to Cairo in the spring of 2011, a few months after the fall of President Hosni Mubarak. Egypt was unsettled but jubilant, and the rest of the Middle East had not yet fallen into war or renewed authoritarianism. One of the Egyptians I interviewed was a blogging pioneer whom I will call Hani.[1] In the early 2000s, Hani had been among the first to take advantage of the burst of freedom experienced by Egyptians before the authorities fully caught on to the Internet's revolutionary potential. Many bloggers made it through the Mubarak era largely unscathed because the government could not keep up with or fully understand the new medium. Unfortunately, Hani had caught the attention of the government; he was tried and sentenced to years in prison for the crime of insulting Mubarak. Throughout his imprisonment, he remained defiant. He was released in November 2010, just months before a Facebook page would spark a revolution that would dramatically change the country, the region, and the world.

Before going to jail, Hani felt that his blog had been a bustling crossroads of discussion. His voice reached

doi:10.1162/DAED_a_00366

farther than he had ever thought possible. After his multiyear involuntary hiatus, he resumed blogging, he told me, with enthusiasm and excitement. But when he came out of jail in late 2010, he found that his blog, and much of the Egyptian blogosphere, had become a comparative wasteland.

"Where is everybody?" Hani answered himself: "They're on Facebook."

At the time I interviewed him, that did not seem like such a bad development. Just a few months earlier, a Facebook group titled "We are All Khaled Saed" – named after a young Egyptian man who had been tortured and killed by the police – had become the organizational core of the revolution. The page was created in June 2010, a few days after Saed's death became public knowledge. The (then-anonymous) administrator of the page was Wael Ghonim, a Google employee and early adopter of the Internet in the region. Ghonim had foreseen Facebook's potential to reach large numbers of ordinary people: in just one month, his page gathered more than one hundred thousand readers, and ordinary Egyptians began using it to engage in political discussion.[2] In later interviews, some of those who participated on the page told me that they felt jubilant and liberated to be finally speaking about politics with other Egyptians.

After the Tunisian revolution of early 2011, the "We are All Khaled Said" page became a hotbed of conversation for Egyptians who longed for a similar upheaval. After much discussion, including polls asking the page's readers what they thought should be done, Wael Ghonim created an event titled "The Revolution," scheduled for January 25, 2011, which was already a traditonal day of protest in Egypt.[3] Hundreds of thousands of Egyptians accepted an "evite" to "The Revolution," displaying their dissent openly, many perhaps for the first time, to their online social networks.

Emboldened by the outpouring of dissent, thousands of people assembled in Tahrir Square on January 25, 2011. One year prior, only about one hundred protesters met in Tahrir Square, where they were surrounded and outnumbered by the police. But this year, the protest quickly swelled to include hundreds of thousands of Egyptians who occupied the square until Mubarak stepped down. To many activists I talked with, Facebook's reach felt empowering. A survey of Tahrir protesters confirmed that social media had been essential to the early turnout that had triggered the avalanche of dissent.[4] Egyptian use of Facebook continued to grow, and it became plainly obvious that Facebook had become a major player in the civic sphere. Even the new military council that replaced Mubarak launched a Facebook page.

But what did it mean for Facebook, a corporate platform, to become so central to the political life of the country? That was less clear.

With the advent of social media platforms in the mid-2000s, the "networked public sphere" – the burgeoning civic space online[5] that had been developed mostly through blogs – expanded greatly, but with a simultaneous shift to commercial spaces.[6] Many scholars and civic activists worried about how "sovereigns of cyberspace," as Internet-freedom advocate, journalist, and author Rebecca MacKinnon called these online platforms, would wield their power.[7] Would they censor and restrict freedoms to serve advertisers or governments with whom they were trying to curry favor? Would they turn over user information to repressive regimes? MacKinnon was prescient in identifying the core problem: the growth of privately owned spaces that functioned as public commons. Over time, the threats posed by this relationship may exceed even our earlier concerns about censorship.

Driven by structural dynamics and corporate motivations, as well as by characteristics of the Internet, these new social platforms are remaking the Internet in a way

that imperils the open architecture of the early Web that felt so intoxicatingly empowering to many of its users. The consequences are profound. This article examines where we are now, and then briefly traces the dynamics that have led us here.

In 2015, Hossein Derakhshan – who has been called the "grandfather" of the Iranian blogosphere – left prison after serving six years of a nineteen-year sentence for blogging, including long stretches of solitary confinement. But prison did not break him; instead, he says, what nearly broke his heart was what he found online when he started blogging again.[8]

After being released, Derakhshan learned that he needed to adapt to the new digital environment and use the new commercial social networks. Up for innovation and change, he created a Facebook account and posted a link to his blog. To his dismay, his post disappeared after just a few "likes." Likes are the main currency in Facebook's all-important algorithm that decides which posts to display to other users, and which to hide. In the new world of social media, posts like Derakhshan's could disappear without being seen by more than a handful of people. Derakhshan was despondent about trying to learn the ropes of this new world. But he soon realized that his personal grasp of the platform was not the only missing ingredient.

The new platforms were strangling access to the hyperlink, directing users to content within their walls and regulating access to the outside Web in very specific ways. Content like his, which was hosted outside of Facebook's territory, did not stand a chance.

Derakhshan wrote the essay "The Web We Have to Save" about his new experience of being online:

> Nearly every social network now treats a link as just the same as it treats any other object – the same as a photo, or a piece of text – instead of seeing it as a way to make that text richer. You're encouraged to post one single hyperlink and expose it to a quasi-democratic process of liking and plussing and hearting: Adding several links to a piece of text is usually not allowed. Hyperlinks are objectivized, isolated, stripped of their powers.
>
> At the same time, these social networks tend to treat native text and pictures – things that are directly posted to them – with a lot more respect than those that reside on outside web pages. . . . A link to the pictures somewhere outside Facebook . . . are much less visible to Facebook itself, and therefore get far fewer likes. The cycle reinforces itself . . . Instagram – owned by Facebook – doesn't allow its audiences to leave whatsoever. You can put up a web address alongside your photos, but it won't go anywhere. Lots of people start their daily online routine in these cul de sacs of social media, and their journeys end there.[9]

There are billions of people on the Internet, but a few services capture or shape most of their activities. Take Facebook: it has 1.5 billion users, a billion of whom log in daily to see updates and news from the hundreds of people they have "friended" on the platform.[10] Or consider Google: more than one billion people use the site to run more than three billion Google searches per day. Facebook recently announced a program encouraging publishers to upload articles to Facebook's servers to make them appear faster to the end-users. Google is planning a similar gambit with "instant" articles of its own. As smartphones continue to claim an increasingly large share of Internet users, Google is also designing a new way to display pages on mobile devices.[11] Google's new scheme would shift more power to the company; though, as with all the other transitions, it would offer benefits to users as well, which often serve to mask, or at least make palatable, the expansion of power.

For an increasing number of people, Facebook and Google are *the* Internet, or at

least the framework that shapes their experience of it.[12] These platforms own the most valuable troves of user data; control the user experience; have the power to decide winners and losers, through small changes to their policies and algorithms, in a variety of categories, including news, products, and books; and use their vast earnings to buy up potential competitors.

I talked with Derakhshan (online, since he is still in Iran) about his experiences, sharing my own research about the shift to a world of algorithmic walled gardens. Both of us are aware that current social media platforms reach many more people than the Internet did in the heydays of blogging. That is not the problem. Neither is it the existence of more frivolous or mundane content online; cute cat and baby images are part of the package. The problem is the shift in the architecture of the Internet. In ways both dramatic and subtle, the shift has begun to create new profound and far-reaching problems. In Derakhshan's words, a link is not just a link; it is a relationship. The power of the Internet comes from our relationships on it. And these relationships are increasingly mediated by the platforms that collect data about us; make judgments about what is relevant, important, and visible; and seek to shape our experiences for commercial or political gain.

How did we get here? And how much power is now concentrated in these platforms? The answers to these questions are connected and offer hints of possible alternative futures.

Legal scholar Lawrence Lessig has famously listed four forces that shape "cyberspace": law, norms, markets, and code.[13] He compared his model to the offline world where law, norms, markets, and architecture play a major role in shaping society. Lessig analogized computer code, which defines how online platforms work, to the role architecture plays offline. Take the layout of a city, for example: When residential and office buildings are separate, and people live in far-flung suburbs, there are social, political, and cultural consequences. Low walkability may contribute to unhealthy lifestyles. Or political polarization may increase while people segregate by income levels and race.

Online, computer code offers a similar structuring power. For example, Facebook requires mutual consent to interact, while Twitter allows people to "follow" someone else without being followed back. On Facebook, friending someone requires acquiescence on both sides: the person making the request and the person accepting it. On Twitter, any public account can be followed with just a click, without having to formally ask for permission. These structures are formed through decisions made by the people who run, administer, and create the code for these platforms, and are implemented by in-house coders, resulting in different social and political environments for each service. Facebook tends to have smaller networks made up of friends, family, and acquaintances, while Twitter is better suited for fan/celebrity relationships in which the few can be followed by the many. Online platforms are shaped not only by the code that structures visibility and access, but by computation and data as well. This combination gives online platforms powers for which there are no simple analogies in the offline world.

The massive accumulation of user data has been written about extensively.[14] There is an increasing amount of data about everyone. More and more social, political, and financial interactions are performed online. More and more people carry phones that connect to the Internet and log their location and activities. Everyday objects are increasingly acquiring sensors that collect information even about passersby. Some of these data are accessed by governments for political purposes; some are used by com-

panies and advertisers for marketing. Financial institutions mine data to check credit-worthiness. Occasionally, the data are leaked, hacked, or otherwise released for reasons that can range from crime to politics to mischief. Ordinary people have very little idea about who holds what kind of data about them, or how the data are used. The amount of accumulated data and the asymmetry of power between the people who are monitored and surveilled and the platforms in which the data are held and mobilized is a significant problem, confirmed by polls revealing the public's great uneasiness about surveillance.[15]

However, the involuntary accretion of massive amounts of data about people is only the tip of the iceberg. In a networked society, computation brings another dimension of asymmetric power. Through techniques that can be loosely collected under the heading "computational inference" – the application of statistical methods, modeling, and machine learning to vast troves of data to make predictions – those who have gathered these data can infer from them information that has never even been disclosed.[16]

In other words, aided by computation, big data can now answer questions that have never been asked about individuals who are the sources of the data:

> The advent of big datasets that contain imprints of actual behavior and social network information – social interactions, conversations, friendship networks, history of reading and commenting on a variety of platforms – along with advances in computational techniques means that political campaigns (and indeed, advertisers, corporations and others with the access to these databases as well as technical resources) can model individual voter preferences and attributes at a high level of precision, and crucially, often without asking the voter a single direct question. Strikingly, the results of such models may match the quality of the answers that were only extractable via direct questions, and far exceed the scope of information that could be gathered about a voter via traditional methods.[17]

The computational inference generated by machine learning takes place during the process of sifting through many varieties of data, with the proviso that the data are deep and rich enough. Inferring political variables about a person does not require their participation in overtly political websites or conversations. For example, Facebook operates mainly through likes: a one-click operation that signals a user's approval of a page, update, or person. The collection of these likes can be used to model, with surprisingly high statistical reliability, a range of outcomes, including "sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender."[18]

This type of analytic power can go beyond many of the traditional categories used by demographers and advertisers to profile the public. By using only their social media imprints (again, not directly asking questions of individuals), researchers have been able to identify people who are likely to become clinically depressed in the future, even *before* the onset of clinical symptoms.[19] Much of this research is done with the best of intentions: for example, as early intervention for new mothers at risk for postpartum depression.[20] However, it is easy to see the downsides of making inferences using data in this fashion. Advertisers, for example, discovered that when women feel "lonely, fat, and depressed" they are more likely to purchase makeup, and that such women are ideal targets for "beauty interventions."[21] In other words, women who are depressed and lonely can be more easily sold makeup. It does not take much imagination to see that advertisers will therefore want to use data gathered by online platforms to find out exactly who is feeling "lonely, fat, and depressed" and mar-

ket to these targeted women at exactly these times.

The increasing use of opaque computational methods known as machine learning – or "neural networks" – adds another layer of complexity to predictions made with big data sets. These are systems that "learn to learn" how to classify individuals (or whatever type of cases they are presented with) into various categories. Machine-learning systems are often provided with a "training set": a database in which cases are marked with the correct answers.

For example, to train a machine-learning system, an employer might provide it with a list of employees he has classified as either "high-performance" or "unsatisfactory," accompanied by social-media data about all employees in the database. Without receiving direct instruction or a recipe about what makes a worker either high-performance or unsatisfactory, the system learns the set of associations that are linked to each outcome, and how to use that knowledge to classify new employees. On the surface, this looks a lot like many other methods that employers use to discriminate among potential hires. But there is a twist: a machine-learning system often does not provide any human-understandable clues to why it classifies the way it does. In fact, if we knew exactly what it was doing, there would be no need for the "machine-learning" part: we could just program the criteria ourselves. In reality, though, all that a manager might know is that the system places potential hires into one category or the other, without any understanding of what parts of the social media big data set were used as signals for a particular outcome.

For all a hiring manager knows, such a system might classify applicants based on criteria such as statistical likelihood of experiencing depression in the future (even if undiagnosed at the time of evaluation) or the possibility of impending parenthood. It is well known that current hiring systems incorporate many biases. However, if we use social media data churned through computational methods for hiring, we may move from imperfect hiring systems that we know discriminate against women, for example, to ones whose workings are hidden from us, but nonetheless still discriminate. This could mean using systems that discriminate *only* against women who are statistically likely to become pregnant soon. This type of discrimination would not be visible to employers because neither the women being hired nor the women not being hired would be pregnant at the time of the hiring, and because a machine-learning system does not display decision-making variables that are easily interpretable, even by its engineers. Social media platforms increasingly hold the kind of data that can be used in these ways.

While this combination of big data and computation obviously creates significant challenges, there are additional, equally daunting issues. When combined with the power of "code" as architecture, in the sense first identified by Lessig,[22] platforms can also nudge behavior, quietly and imperceptibly, and sometimes in ways that are not directly visible even to the people who run the platforms. Facebook, for instance, uses an algorithm to order the news feed that shows its 1.5 billion users' status updates. These may range from updates that are purely personal in nature to news articles. Increasingly, for many population segments ranging from younger people in developed countries to populations just coming online in poorer countries, Facebook has become the number one source of news.[23] In poorer countries, many people are not even aware that there is an Internet outside of Facebook,[24] and many others choose to stay completely within Facebook's realm.[25] As David Clark explains in his essay in this issue, Facebook has helped ensure this through promotion of its stripped-down Facebook

app – 0.facebook.com – which, in agreement with mobile service providers in many developing countries, does not incur data charges for users.[26]

In my research, I have encountered many people whose Internet routine resembles the following: If on a desktop computer, a user launches a browser and types "Facebook" into Google's search box, likely unaware that the URL bar at the top of the browser is a separate and faster way to get there. Google brings up Facebook as the first link, and the user clicks on Facebook and proceeds to interact mostly within the site. If using a mobile platform, which is increasingly the norm, a user will simply launch the Facebook app and rarely encounter the open Web at all.

This tendency to stay within Facebook is what gives Facebook's architectural decisions such power, and invisibly so. In one study, 62.5 percent of users had no idea that the algorithm controlling their feed existed, let alone how it worked.[27] This study used a small sample in the United States, where the subjects were likely more educated about the Internet than many other populations globally, creating a potentially unrepresentatively low estimate. The news feed is a world with its own laws of physics, and the deities that rule it are Facebook programmers. In this world, some types of information are nudged and helped to spread more, while others are discouraged. There is great power in what we do (and do not) see from our friends and acquaintances, and increasingly, this is greatly influenced by platform design and code.

Facebook's own research has shown the power of its designers' architectural choices. In one Facebook experiment, randomly selected users received a neutral message to "go vote," while others, also randomly selected, saw a slightly more social version of the encouragement, noting also which of their friends voted using small thumbnails of their profile photos. Matched with voter rolls, that single message caused 340,000 additional people to turn out to vote in the 2010 U.S. congressional elections.[28] In another experiment, Facebook randomly selected whether users saw posts with slightly more upbeat words versus more downbeat ones: the result was correspondingly slightly more upbeat or downbeat posts by those same users. Dubbed the "emotional contagion" study, this incident sparked international interest in Facebook's power to shape the user's experience.[29]

The power to shape experience (or perhaps elections) is not limited to Facebook; there are other powerful platforms. For example, Google rankings are hugely consequential. A politician can be greatly helped or greatly hurt if Google chooses to highlight or hide, say, a link to a corruption scandal on the first page of its results. A recent study showed that slight changes to search rankings can shift the voting preferences of undecided voters, and that these shifts can be hidden so that people show no awareness of the manipulation.[30]

For a small taste of how platform choices affect the civic sphere, consider the case of the protests in Ferguson, Missouri, in August 2014. What started as a community shaken over the police killing of a young man under murky circumstances grew into major protests after the police responded to initial small-scale – and completely nonviolent, according to journalists on the scene – protests by residents with overwhelming force, including the use of attack dogs and tear gas. A few national journalists, as well as ordinary citizens with smartphones, started tweeting from the scene of the initial protests. The burgeoning unrest and conflict soon grew into major Twitter discussions that later sparked the attention of the mainstream news media. About three million tweets were sent before the mass media began covering events in Ferguson. The nationwide movement that grew from these events is often referred to as the "Black Lives

Matter" movement, named after the Twitter hashtag.

However, on the first night of the protests, the topic was mostly invisible on Facebook's algorithmically controlled news feed.[31] Instead, the "ice bucket challenge," in which people poured cold buckets of water over their heads and, in some cases, donated to an ALS charity, dominated the Facebook news feed. This was not a situation that reflected Facebook users' lack of interest in the Ferguson protests; rather, it was an indication that it is hard to "like" – Facebook's dominant algorithmic signal – such disturbing news, while it is easy to give a thumbs-up to a charity drive. Once a topic is buried by an algorithm, this becomes a self-feeding cycle: fewer people are able to see it in the first place, with fewer still choosing to share it further, causing the algorithm to bury it deeper. On Twitter's platform, in which users see all posts from the people they follow in chronological order, the topic grew to dominate discussion, trending locally, nationally, and globally, catching attention of journalists and broader publics. On Facebook, it barely surfaced. Given the importance of online platforms and public attention to political movements, burying such news is highly consequential.[32] Had our media been exclusively controlled by an algorithm in which "liking" were the main emotive input, the long and hard national conversation about race and policing in America that was generated by the Ferguson protests might have never transformed into a national movement.[33]

How did we get here? Was it inevitable? Tracing this path requires combining and probing the two questions posed by Hani and Derakhshan, two people who blogged under repressive regimes and who were released from prison five years apart. Why is everyone on Facebook now? And why are these platforms killing the hyperlink as an independent relationship between people? Why are they dictating who sees what?

Some aspects of the answer are deceptively simple, and at the same time deeply structural. The open Internet that held so much generative power took a turn toward ad-financed platforms, while the dangers lurking for ordinary users from the Internet's open and trusting design were not counteracted, causing people to flee to safer and more user-friendly platforms. In combination, these two developments encouraged, enabled, and forced the creation of massive, quasi-monopolistic platforms, while incentivizing the platforms to use their massive troves of data with the power of computational inference to become better spy machines, geared toward ad delivery, the source of their financing.

From Wikipedia to question-and-answer sites to countless numbers of sites and blogs that provide a public service (but not payment for their creators), the Internet offers direct proof that people enjoy sharing their creative and personal output with others.[34] If there were ever a need to expand our conception of humanity beyond the restricted "homo economicus" who works only for his or her benefit, the explosion of user-generated content on the Internet has provided major evidence.[35] However, creative and altruistic output alone does not provide financing for servers, coders, and database management. As the public Internet scaled up and grew in numbers of participants, many websites faced a dilemma: whether to charge their users, or to sell users' eyeballs to advertisers.

It was a crucial turning point: were people going to be the customers, or were they going to be the product sold? Almost all of the major platforms went with advertising. As Ethan Zuckerman, then a staff member of one of the Internet's earliest user-generated platforms, tripod.com, explains:

> Advertising became the default business model on the web, "the entire economic founda-

> tion of our industry," because it was the easiest model for a web startup to implement, and the easiest to market to investors. Web startups could contract their revenue growth to an ad network and focus on building an audience. If revenues were insufficient to cover the costs of providing the content or service, it didn't matter – what mattered was audience growth, as a site with tens of millions of loyal users would surely find a way to generate revenue.[36]

These decisions were made partly out of idealism: a free website-hosting platform like Tripod also allowed Thai dissidents to circulate otherwise censored content without worrying about paying for the site. It made more sense at the time to have ads than to charge users. But once advertising became the way to make money, almost everything flowed from it, especially when combined with another key feature of online platforms: network effects.

Network effects, also called network externalities, are the tendency of the value of some products or services to increase as more people use them, and to become less worthwhile when they are not used by others, even if the less popular product or service is objectively better, cheaper, faster, or more diverse in its offerings. For many online applications, everyone wants to be where everyone else is. This dynamic allows many online platforms that manage to get ahead of their competition to completely dominate their niche:

> The more people own fax machines, for example, the more useful each one becomes. That is also why there is a single standard for fax machines – would you switch to a brand new, faster fax machine standard if there was nobody else you could fax with your machine? Research shows that the presence of network externalities trumps product preference or quality; many people will chose a service that has more users compared to the one that is otherwise better for them. Such platforms, such as Facebook, tend to quickly dominate their market and become near-monopolies. This is also why everyone lists their wares on Ebay, where all the buyers are, and advertises on Google, where all the eyeballs go. The fact that a lot of people already have Facebook accounts means that considerations of network externalities will result in existing people staying put, or new people joining in anyway, even if they have qualms about the privacy issues.[37]

While network externalities made it possible for platforms to become very large, the ad-financing model meant that a mid-sized platform, even one with hundreds of millions of users, faced great challenges, since ads on the Internet are not worth much.[38] An ad-dependent platform can only survive if it serves enormous numbers of people. For example, Wall Street's investors have soured on Twitter because it only has about three hundred million users. For most products, hundreds of millions of users would appear to be a huge success. In an ad-financed online world, that's barely enough to get by.

But there is one key path for online ads to become more valuable for platforms. If platforms accumulate a great amount of data on their users, and harness computational inference to "understand" them on behalf of their advertisers, then the ads, which have a higher chance of leading to a purchase, are worth a lot more. These advertisers could include both corporate entities selling products and political campaigns marketing politicians. Platforms can also use their architectural power to create an environment that is more advertiser-friendly. Until quite recently, for example, Facebook allowed likes as the only signal (aside from making comments) that users could send about a page or status update. While Facebook recently expanded choices in a few countries to include a few more "one-click" options such as "like," "love," "haha," "yay," "wow," "sad," and "angry," the

expanded list is still heavily geared toward positivity, with only two that are typically associated with negativity: *angry* and *sad*.

Overall, many of the issues identified in this article are a direct consequence of this combination: Internet platforms are financed by ads that demand great scale, and they are fueled by network effects that allow such scale through the emergence of monopolies. These quasi monopolies then have incentives to collect and process vast amounts of data on their users to make the ads more effective for the advertisers, while also controlling the experience of the users to keep the platform advertising-friendly, and to keep the user from leaving the platform.

The other major development over the past decade from the user side has been the lack of attention and resources to ensure that the open Web – the one in which the hyperlink and address bar, rather than a closed platform and its algorithmic and architectural choices, dominate navigation – remains a secure and navigable place for ordinary users.

Many of the early protocols that defined the Internet were developed for use by a trusting, small, and closed community of academic and military research staff. However, on the current scale of billions of people, the Internet's insecurity, and the proliferation of malware, spam, and untrustworthy sites, has caused many to retreat to easier-to-use, relatively safe platforms. The ad-financing model means that almost all commercial websites have installed extensive ad-tracking software on their sites, which is not distinguishable, in effects or operation, from malware dedicated to spying. Navigating the ordinary, open Internet now seems treacherous and feels slow (since the sites are loaded with ads and tracking software).

In 2014, for example, a massive vulnerability was found in "OpenSSL," one of the protocols that underpins almost all Internet commerce. The bug "heartbleed" allowed an attacker to read parts of a computer's memory that the program should not ordinarily have access to, and to learn crucial private information, including stored passwords. While it is almost too ridiculous to believe, the OpenSSL architecture, used by about two-thirds of all web pages, including almost all major banks, is maintained by a group of only a dozen people, all but one of whom are volunteers.[39] The crisis with OpenSSL was but one example of critical parts of the Internet's infrastructure that provide security for ordinary users being tended by almost nobody. There is very little energy or resources dedicated to tending the commons of the Internet, and the resulting environment has made ordinary Web navigation increasingly difficult and user data increasingly insecure. For regular users, remaining within trusted walled gardens, like Facebook or Google's new proposed Web architecture, is a reasonable choice. This is exactly the scenario warned against by scholars.

This shift toward the walled gardens is only increasing as the next billions come online: people with less technical literacy, less powerful devices, shakier Internet connections, and often mobile-only access. In developing nations, the walled gardens of huge online platforms have many draws. Network effects means that their expatriate relatives and friends are most likely to be on the biggest platforms. A controlled environment makes the Internet more navigable. Bigger platforms offer better translation and localization services, something volunteer sites have more difficulty providing. Google helps order the chaotic, seemingly endless, choices effectively, while Facebook offers a way to manage the flow of information from a user's social networks, albeit algorithmically curated within an ad-delivery platform. And thus, the Internet giants continue to grow, and have become

the dominant landscape of the Internet for most people.

In his prescient book *The Future of the Internet – And How to Stop It*, Jonathan Zittrain warned about these problems, and predicted that unless addressed, they would lead to the collapse of the open, generative Internet in favor of closed systems.[40] Legal scholar Tim Wu looked into past information systems and pointed out that many became dominated by monopolies.[41] As early as 2003, Deborah L. Spar – now president of Barnard College – predicted that insurgent technologies would pass from the "pirates" that use technologies to disrupt order to the hands of powerful commercial and governmental bodies who use it to consolidate power.[42] The Internet, in some ways, seems set on this path, although we have not yet passed the point of no return.

Because ad-based financing quickly devolves into large-scale, monopolistic systems working on behalf of advertisers, to change directions, we first must change how we finance the Internet's platforms, including financing potential challengers to currently dominant platforms. Alternative models of financing were developed in the early days of the Internet, but these were quashed, in part because they may have been too early for mass adoption, but also because banks and websites resisted their implementation. Second, the Internet's commons needs tending, which will also require substantial resources and financing as well. A global system whose security depends so much on volunteer work will, inevitably, become a difficult-to-navigate, insecure, and unpleasant experience, and will result in people retreating to safer platforms that cushion the user experience while also controlling it. Third, we must recognize that due to network effects, unregulated markets (one of the mechanisms of Lessig's original four forces) do not work well on the Internet for certain kinds of platforms, including many of the current tech giants. The influence of network effects is especially powerful for user-generated platforms, since what partially powers them is user investment. People have spent a lot of time and effort building up their positive feedback on eBay and cultivating their social networks on Facebook. It is unlikely that competition alone – even competition offered with better terms and services – can dislodge these powerful platforms, given the costs sunk into them by their users.

The path toward change is uphill, but the first step requires the public recognition of what dissidents in repressive regimes – often the canaries in mines – have already discovered: the power of the Internet derives from our ability to freely connect with each other. These developments are not changing only from one type of program or site to another; they are shifting to a new regime in which our relationships are mediated by forces trying to mine our data, mostly in order to sell a few more ads slightly more effectively, but also open to a host of other political uses.[43] From politics to culture, much power resides with owners of data, especially those possessing command of computation and online architecture.

It is not too late to change this path, but to do so requires an open-eyed and realistic look at the forces that have brought us here – financing models, the need for tending the security of the Internet's commons, demand for usability, and the shift to mobile – and asking how to generate an alternative model that can scale-up. That demand still exists: the first billion Internet users have experienced, and remember, the admittedly chaotic early Internet, built upon the energy and euphoria of people discovering both information and each other. Now that the Internet is approaching three billion users, the question facing us is whether their Internet experience will much differ from a tightly regulated coffeehouse within a gigantic shopping mall.

ENDNOTES

1 He did not ask me to keep his identity secret, but I am not using his name on principle, to avoid my arguments getting tangled with his views as a result of Google searches run by clumsy repressive regimes.

2 Jennifer Preston, "Movement Began With Outrage and a Facebook Page That Gave It an Outlet," *The New York Times*, February 5, 2011, http://www.nytimes.com/2011/02/06/world/middleeast/06face.html.

3 Wael Ghonim, *Revolution 2.0: The Power of the People is Greater Than the People in Power – A Memoir* (Boston: Houghton Mifflin Harcourt, 2012); and the author's private conversation with Wael Ghonim (2015).

4 Zeynep Tufekci and Christopher Wilson, "Social Media and the Decision to Participate in Political Protest: Observations From Tahrir Square," *Journal of Communication* 62 (2) (2012): 363 – 379, http://doi.org/10.1111/j.1460-2466.2012.01629.x.

5 Yochai Benkler, *The Wealth of Networks: How Social Production Transforms Markets and Freedom* (New Haven, Conn.: Yale University Press, 2007).

6 Steven Johnson, "Can Anything Take Down the Facebook Juggernaut?" *Wired*, May 16, 2012, http://www.wired.com/2012/05/mf_facebook/.

7 Rebecca MacKinnon, *Consent of the Networked: The Worldwide Struggle for Internet Freedom* (New York: Basic Books, 2012).

8 Hossein Derakhshan, "The Web We Have to Save: The Rich, Diverse, Free Web that I Loved – and Spent Years in an Iranian Jail for – is Dying. Why is Nobody Stopping It?" July 2014, https://medium.com/matter/the-web-we-have-to-save-2eb1fe15a426.

9 Ibid.

10 Don Clark and Robert McMillan, "Facebook, Amazon and Other Tech Giants Tighten Grip on Internet Economy," *The Wall Street Journal*, November 5, 2015, http://www.wsj.com/articles/giants-tighten-grip-on-internet-economy-1446771732.

11 Joshua Benton, "Get AMP'd: Here's What Publishers Need to Know about Google's New Plan to Speed Up Your Website," *Nieman Lab*, October 7, 2015, http://www.niemanlab.org/2015/10/get-ampd-heres-what-publishers-need-to-know-about-googles-new-plan-to-speed-up-your-website/.

12 Leo Mirani, "Millions of Facebook Users have No Idea They're Using the Internet," *Quartz*, February 9, 2015, http://qz.com/333313/milliions-of-facebook-users-have-no-idea-theyre-using-the-internet/.

13 Lawrence Lessig, *Code: And Other Laws of Cyberspace, Version 2.0* (New York: Basic Books, 2006).

14 See, for example, Viktor Mayer-Schönberger and Kenneth Cukier, *Big Data: A Revolution that Will Transform How We Live, Work, and Think* (Boston: Houghton Mifflin Harcourt, 2013).

15 George Gao, "What Americans Think about NSA Surveillance, National Security and Privacy," Pew Research Center, May 29, 2015, http://www.pewresearch.org/fact-tank/2015/05/29/what-americans-think-about-nsa-surveillance-national-security-and-privacy/.

16 Zeynep Tufekci, "Engineering the Public: Big Data, Surveillance and Computational Politics," *First Monday* 19 (7) (2014), http://dx.doi.org/10.5210/fm.v19i7.4901.

17 Ibid.

18 Michal Kosinski, David Stillwell, and Thore Graepel, "Private Traits and Attributes are Predictable from Digital Records of Human Behavior," *Proceedings of the National Academy of Sciences* 110 (15) (2013): 5802 – 5805, http://doi.org/10.1073/pnas.1218772110.

19 Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz, "Predicting Depression via Social Media," in *Proceedings of the Seventh International AAAI Conference on Weblogs and*

Social Media (Palo Alto, Calif.: Association for the Advancement of Artificial Intelligence, 2013), http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewFile/6124/6351.

20 Munmun De Choudhury, Scott Counts, Eric Horvitz, and Aaron Hoff, "Characterizing and Predicting Postpartum Depression from Shared Facebook Data," in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (New York: Association for Computing Machinery, 2014), 626 – 638, http://doi.org/10.1145/2531602.2531675.

21 Lucia Moses, "Marketers Should Take Note of When Women Feel Least Attractive," *AdWeek*, October 2, 2013, http://www.adweek.com/news/advertising-branding/marketers-should-take-note-when-women-feel-least-attractive-152753.

22 Lessig, *Code: And Other Laws of Cyberspace, Version* 2.0.

23 Amy Mitchell, Jeffrey Gottfried, and Katerina Eva Matsa, "Millennials and Political News," Pew Research Center, June 1, 2015, http://www.journalism.org/2015/06/01/millennials-political-news/.

24 Mirani, "Millions of Facebook Users have No Idea They're Using the Internet."

25 World Wide Web Foundation, *Women's Rights Online: Translating Access into Empowerment* (Washington, D.C.: World Wide Web Foundation, 2015), http://webfoundation.org/wp-content/uploads/2015/10/WomensRightsOnlineWF_Oct2015.pdf.

26 See David D. Clark, "The Contingent Internet," *Dædalus* 145 (1) (Winter 2016), 9 – 17.

27 Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig, "'I Always Assumed That I Wasn't Really That Close to [Her]': Reasoning about Invisible Algorithms in the News Feed," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (New York: Association for Computing Machinery, 2015), 153 – 162, http://www.researchgate.net/profile/Kevin_Hamilton/publication/275353888__I_always_assumed_that_I_wasn't_really_that_close_to_her___Reasoning_about_Invisible_Algorithms_in_News_Feeds/links/553aa2fd0cf245bdd764475f.pdf.

28 Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler, "A 61-Million-Person Experiment in Social Influence and Political Mobilization," *Nature* 489 (7415) (2012): 295 – 298, http://doi.org/10.1038/nature11421; and Jonathan Zittrain, "Facebook Could Decide an Election Without Anyone Ever Finding Out," *The New Republic*, June 1, 2014, https://newrepublic.com/article/117878/information-fiduciary-solution-facebook-digital-gerrymandering.

29 Lorenzo Coviello, Yunkyu Sohn, Adam D. I. Kramer, Cameron Marlow, Massimo Franceschetti, Nicholas A. Christakis, and James H. Fowler, "Detecting Emotional Contagion in Massive Social Networks," *PLoS ONE* 9 (3) (2014): e90315, http://doi.org/10.1371/journal.pone.0090315.

30 Robert Epstein and Ronald E. Robertson, "The Search Engine Manipulation Effect (SEME) and Its Possible Impact on the Outcomes of Elections," *Proceedings of the National Academy of Sciences* 112 (33) (2015): E4512 – E4521, http://doi.org/10.1073/pnas.1419828112.

31 Zeynep Tufekci, "The Medium and the Movement: Digital Tools, Social Movement Politics, and the End of the Free Rider Problem," *Policy & Internet* 6 (2) (2014): 202 – 208, http://doi.org/10.1002/1944-2866.POI362.

32 Zeynep Tufekci, "Algorithmic Harms beyond Facebook and Google: Emergent Challenges of Computational Agency," *Colorado Technology Law Journal* [formerly *Journal on Telecommunications and High Technology Law*] 13 (2) (2015): 203 – 218; and Zeynep Tufekci and Deen Freelon, "Introduction to the Special Issue on New Media and Social Unrest," *American Behavioural Scientist* 57 (7) (2013): 843 – 847.

33 Tufekci, "Algorithmic Harms beyond Facebook and Google."

34 Benkler, *The Wealth of Networks*.

35 Yochai Benkler, *The Penguin and the Leviathan: How Cooperation Triumphs over Self-Interest* (New York: Crown Publishing Group, 2011).

36 Ethan Zuckerman, "The Internet's Original Sin," *The Atlantic*, August 14, 2014, http://www.theatlantic.com/technology/archive/2014/08/advertising-is-the-internets-original-sin/376041/.

37 Zeynep Tufekci, "Facebook, Network Externalities, Regulation," Technosociology, May 26, 2010, http://technosociology.org/?p=137.

38 Zeynep Tufekci, "Mark Zuckerberg, Let Me Pay for Facebook," *The New York Times*, June 4, 2015, http://www.nytimes.com/2015/06/04/opinion/zeynep-tufekci-mark-zuckerberg-let-me-pay-for-facebook.html.

39 Dan Goodin, "Critical Crypto Bug in OpenSSL Opens Two-Thirds of the Web to Eavesdropping," Ars Technica, April 7, 2014, http://arstechnica.com/security/2014/04/critical-crypto-bug-in-openssl-opens-two-thirds-of-the-web-to-eavesdropping/; and Jose Pagliery, "Your Internet Security Relies on a Few Volunteers," CNN Money, April 18, 2014, http://money.cnn.com/2014/04/18/technology/security/heartbleed-volunteers/index.html.

40 See Jonathan Zittrain, *The Future of the Internet – And How to Stop It* (New Haven, Conn.: Yale University Press, 2008).

41 Tim Wu, *The Master Switch: The Rise and Fall of Information Empires* (New York: Knopf Doubleday Publishing Group, 2010).

42 Debora L. Spar, *Ruling the Waves: From the Compass to the Internet, a History of Business and Politics along the Technological Frontier* (New York: Mariner Books, 2003).

43 Tufekci, "Engineering the Public"; and Zittrain, "Facebook Could Decide an Election Without Anyone Ever Finding Out."

Design Choices for Libraries in the Digital-Plus Era

John Palfrey

Abstract: Libraries are more important, not less so, in a digitally networked era. Despite the fact that today's mobile devices feature Google's search box and Apple's Siri to help us find a quick answer to just about any question, we ought to be investing more capital than ever in our public libraries. We need libraries in the digital era to provide a public option to ensure sustained, free, equitable access to knowledge and preservation of our cultural and scientific heritage. In a period when both the analog and digital are useful, the design choices for those building, and reimagining, libraries are many and complex. We ought to design our libraries to meet the near-term possibilities of a networked environment, as well as the long-term requirements of democratic societies and the practice of scholarship. These design choices involve trade-offs and new commitments that may pit future activities against entrenched present-day interests. The essential design choice is between reliance on ever-more efficient interfaces, often developed by commercial outfits, and interfaces that are developed by the library community, engaging the public in coproduction and extending outward via the networked public sphere. The fate of libraries as vibrant institutions with broad public support could turn on the outcome of these design decisions. The challenges facing libraries also inform conversations about the future of other public-facing institutions, such as schools and newspapers, which are important contributors to an informed citizenry and a vital republic.

JOHN PALFREY is Head of School at Phillips Academy in Andover, Massachusetts. Previously, he was the Henry N. Ess III Professor of Law and Vice Dean for Library and Information Resources at Harvard Law School. He is the author of *BiblioTech: Why Libraries Matter More Than Ever in the Age of Google* (2015).

The main building of the Chicago Public Library (CPL) occupies a full city block downtown. From the outside, the building is massive and imposing, yet also appealing in an institutional way. Once inside, however, the building is far from intuitive. You are not met by a warm and welcoming reading room. There's not an obvious pathway to the popular books and DVDs, which are floors away. You find yourself instead in a warren of long hallways, occasionally punctuated by guards and metal detectors: unmistakable signs of the time we live in and the realities of running a public institution in a big city.

The essential concern that animates this essay is this: what if people turn away from imposing buildings like the massive Chicago Public Library and turn instead to their mobile devices, serviced by commercial firms, to meet their needs for knowledge and information in an increasingly digital future? If that

doi:10.1162/DAED_a_00367

comes to pass, what will democratic societies lose? And can librarians, or any of us, do something to head off this outcome?

Brian Bannon, the commissioner of the Chicago Public Library system, is responsible for keeping his library essential and relevant to his city. Bannon and his team know they have to make the main library building, as well as its dozens of branch libraries, accessible and inviting to today's diverse population of Chicagoans. More broadly, Bannon has the job of ensuring that his library is a relevant public institution in a digital age. Libraries cannot afford simply to continue doing things as they have done in the past, assuming that the public will find them useful and that public officials will, in turn, continue to fund them. Bannon – not alone among big-city library leaders, but among the most inventive and forward-looking – has acknowledged this challenge and is meeting it head on, with a modern design sensibility and a highly networked approach.

The challenge facing all library leaders as we transition from an analog era (of atoms) to a predominantly digital one (of bits) is multifaceted. Materials today are typically created using digital platforms, though they are often later rendered in other formats. Think of this journal, or the newspaper you read in the morning. Each was initially produced as a digital file, on a computer (this essay, using Microsoft Word and Google Docs), and then rendered in a variety of formats. In the case of this issue of *Dædalus*, it will exist as both a printed journal and as a digital file, perhaps even in a range of digital formats. In the future, materials will be mostly accessed in their digital forms; the trends in use clearly point in that direction. It is for this reason that I argue we are in a "digital-plus" era of libraries: not everything must be digital, but materials tend to be born digital and, thereafter, take a variety of forms through which people access them.

The need for libraries to provide information in a heterogeneity of formats stems from a heterogeneity of preferences among those who interact with these materials. The challenge for libraries is to find a way to keep up with the rapid changes in the desires and expectations of citizens regarding the type and nature of information that libraries provide. The formats of materials are shifting rapidly: from vinyl to cassettes to compact discs to MP3 files to streaming services in audio; from enormous film reels to VHS and Betamax to DVDs and Blueray to myriad other digital formats in video and film; from the traditional printed codex (our familiar book format) to all manner of digital files for monographs, journals, and other text-based works. What's more, in each medium there is wide disagreement on which format is superior. And within any community, a librarian finds a split in preferences. While the growth in eBooks is sharp and unmistakable, many readers – of all ages, it turns out – still prefer the feel of a real book in their hands. (Though I work to build digital libraries, I share this same preference.) A librarian needs to be able to meet all these varied desires in order to keep patrons coming back, whether their interests are traditional or new-fangled.

Expectations about the services that librarians ought to provide are also changing quickly. From the libraries of antiquity until well into the twentieth century, a library could succeed by serving as a well-indexed, well-organized storehouse for printed materials. Patrons had no choice but to come to a physical space – often a glorious one – to consult the materials and to seek help in finding knowledge that they had previously not encountered. Today, so much knowledge – as well as misinformation – can be found instantaneously by anyone literate and wealthy enough to own a smartphone, tablet, or computer. Consider what has happened to the market for printed copies of the *Encylopædia Britannica*: is there a single

encyclopedia salesperson going door to door today, as they once did in large numbers? The worlds of reference, of book provision, of news and information – they have all undergone radical change in a few short decades. That change shows no sign of slowing down.

Mr. Bannon faces, as do all library leaders, these many issues at once, without major budget increases to foot the bills. Bannon's answer has been to be both excellent at traditional librarianship and creative about offering new services. He has ensured that his library is a part of broader networks of people working toward common ends. In adjusting the way the Chicago Public Library operates, Bannon has managed to get in front of the changes and align building, staff, and services to the needs and interests of the communities they serve. For instance, one way the Chicago Public Library has attracted more visitors has been to observe what they do in the library and what they do not; to ask them what they want and what they could do without; and to make adjustments accordingly.

One way to preserve libraries as public access points and repositories for culture and knowledge is to ensure that people of all ages keep coming in to use them. That previously uninviting first floor of the main Chicago Public Library building today hosts a large number of adolescents every afternoon in a space called the YouMedia teen learning space.[1] YouMedia is instructive and important on multiple levels. One is that the young people using the space are involved both in the enjoyment of cultural materials and in the creation of new materials. The space includes printed books, attractively set forth on well-positioned shelves, alongside other age-appropriate physical materials. But the space also features a range of digital devices, used both for interacting with and creating digital images, sounds, and text. The staff who work in the YouMedia space are expert at engaging young people in the hybrid world of the digital and the analog. And the YouMedia space is an oasis in the hectic city; whether hot or cold outside, the center is a safe, attractive, warm environment for teens to congregate in, not far off the street, in the city center.

The YouMedia learning space demonstrates the design sensibility that helps keep libraries vibrant, as well as the importance of libraries operating as nodes in a network, rather than as standalone facilities. YouMedia came about as a partnership, not just a library-only activity. The core funding has come from the John D. and Catherine T. MacArthur Foundation; the intellectual capital has come from academics and activists; the staffing has been drawn from multiple networks of people interested in kids, technology, and education, as well as libraries. And in turn, YouMedia facilities have cropped up in other cities (including, for instance, Miami); other foundations have stepped up to fund these related efforts; and a new network, spanning communities around the country, has emerged to support teen learning. One space and one team will always be the first, and surely there is pride of place; but more important, the network effect ensures that the whole is greater than the sum of the parts. Librarians who can create and nurture such partnerships stand to benefit enormously.

To draw in new patrons and build strong networks, the Chicago Public Library has done more than just incubate the YouMedia learning space. Entrepreneurs and tinkerers are drawn into a new "maker" space devoted to innovation and creativity in that same massive building, just upstairs from YouMedia. The Library has also helped to host and support national and international networks of librarians who are involved in the reinvention of libraries, including through the Digital Public Library of America (DPLA) and the international NEXT Library network.

As one of the great urban public library systems in the world, the Chicago Public Library ought to continue to circulate books, operate its eighty appealing branch locations, and manage an important physical collection. In 2014, the roughly thousand-person staff made services available to ten million visitors and circulated another ten million materials.[2] But the CPL takes none of this activity and popularity for granted in what are uncertain times to work in any aspect of the information business. Rather, the CPL is showing how a focus on design sensibility and building networks can ensure that libraries secure both visitor and financial support and thrive into the future.

The design choices for libraries during this period of transition come fast and furious. Each type of library – whether a university academic library, a public elementary school library, a special library, or an archive – must question its priorities and vision for the future. Choices are prompted, in part, by the aging of buildings. More acutely, the changes are brought about by the question of whether to invest more in on-site physical objects, or in digital works. These digital works are not owned and stored in the traditional way – they are not "bought," "shipped," and "shelved" – but are shared and made available to patrons near and far. Librarians who serve patrons directly must decide how to spend their time: by making themselves available at a reference or circulation desk, or by focusing on instant messaging and responding to online queries. Libraries who manage collections must decide whether to emphasize aggregation of shared materials or curation of unique collections.

These library-level design choices roll up to a society-wide decision about access to knowledge. One option is to continue the present trajectory toward access to knowledge increasingly through commercial interfaces. These interfaces are improving in efficiency, beauty, and accessibility at a fast pace. Some of these interfaces are supported through advertising; others are pay-per-use; but uniformly, the leaders are commercial. Libraries, in this model, would accept their reliance on these commercial services and would instead focus their attention inwardly and locally, serving communities through the space the library can provide and by setting these materials in helpful contexts. Another option is for libraries to compete with these commercial firms, developing systems at networked scale that would serve the public through networked interfaces as well as through the local, physical interfaces of their buildings. In this second response, librarians would function as networked actors, their physical presence being nodes in that larger network of public-facing cultural heritage institutions. Either way, librarians must change the way they conceive of their role.

The existing architecture of a wonderful, historic library and shifting formats are not the only constraints facing libraries during the transition from the analog to the digital. The other major challenge involves personnel; put more precisely, the way in which librarians have been trained. Many active librarians learned through countless hours of disciplined practice in a very different environment than now exists, predominantly based in analog materials and associated services. This challenge is not lost on librarians; the topic is addressed, one way or another, at nearly every library conference. A growing cohort of librarians has made the transition from an analog-era outlook to a digital-era approach. Others have not.

Instead of thinking of them as standalone institutions, we ought to reconceptualize libraries as nodes in a network and librarians as networked actors. Each library, or node, serves both people immediately proximate to the library in physical terms and those who are interested in the library's contents but are remote from it. Each librarian,

in turn, is a networked actor, involved in social production of knowledge on a scale much larger than the library in which she or he physically works. For example, in addition to serving patrons standing in front of them, librarians undertake the act of working with a highly distributed group of peers, all devoted to the task of building a digital knowledge commons.[3] The notion of librarianship takes on a yet higher profile: librarians have a role to play, as collaborative actors, in developing, curating, and making accessible the world's knowledge on a grand stage.

Librarians who operate as networked actors are already rethinking their roles and redesigning libraries as institutions, from the inside out. Librarians who learn to hack systems, in collaboration with other public-spirited actors, are positioning themselves for success in their profession as it continues to morph toward the digital. These librarians will focus on serving the public good beyond the immediate needs of the patrons in their community. This shared work will thus serve all communities better.

New platforms can assist networks in functioning well. For libraries and librarians, a common, open, distributed technology platform can bring together the technology, people, code, materials, and spaces in ways that will serve the public during this hybrid era of print and the digital. In the United States, the Digital Public Library of America is designed to serve the role of shared platform. The DPLA is an open distributed system for sharing the cultural, historical, and scientific heritage of the United States. As a platform, it also functions as a test environment or, at the risk of mixing the metaphor, a "sandbox" in which this reinvention process can happen.

On one level, the DPLA is an open-source repository of code, tools, and metadata. As a repository of code, the DPLA makes available the computing know-how of a subset of a community to everyone else in that community. Just as many aspects of the Internet have come to run on open-source code, the library world can share the core systems that make knowledge available broadly. As for tools, the DPLA encourages people to come up with mechanisms for sharing information and knowledge in new, graphic, enticing ways. Examples of these tools include means to integrate library materials with Wikipedia, the peer-produced online encyclopedia; ways to display books and other materials on an online "bookshelf" that can extend forever; and ways to sort materials based on what types of knowledge people have accessed in the past (protecting the identity and privacy of individuals in the process).

As for metadata, the DPLA pulls together the digital materials that librarians from across the globe have digitized and made available online. Large institutions – such as Harvard University, New York Public Library, and the National Archives – contribute millions of records, making them easily accessible to the public. When Harvard digitized its Emily Dickinson papers, for instance, these materials could be accessed directly through the university's websites, or through the DPLA. Likewise, a local library could take the metadata – the data that describe the Dickinson papers – and point directly to the sources for their local users.

Smaller institutions, too, can contribute their materials to this shared repository. Through a series of state hubs, or "on-ramps," to the national database, local historical societies, libraries, and archives can digitize materials and then share them with the world. For instance, a postcard collection held at the Boston Public Library, scanned and shared through this national platform, has made available historical images that show the Little Missouri River in the Badlands of North Dakota and the Ringling Museum in Sarasota, Florida.[4] Maps of the Atlantic continents, showing the change

in perception of land and sea over hundreds of years, held in trust by a New England boarding school, have been digitized and shared such that the maps can be used in classes and by scholars anywhere in the world.[5] These materials would not become accessible to the public without the hard work of librarians, trained in the art of digitization, metadata creation, and online storage.

By operating as nodes in a network, librarians can make their materials more broadly accessible to others, and in turn make accessible to their own users materials held elsewhere, all in an instant. Kept locally, materials are of limited utility; shared globally, via a well-designed network, these materials have far greater value. The DPLA includes large stores of open-access metadata and content, intended for curation and reuse in specific communities.

The DPLA is fundamentally a network of people who want to serve the public as well as possible, by and through digital technologies. The DPLA is designed to be a mechanism to support librarians in crafting and honing their role as nodes in this new, highly networked environment. As a platform, the DPLA can also help support the professional development of librarians (in the service of their patrons). These librarians might still be students; they might be seasoned veterans; they might be retirees with active minds and the will to continue to give back to society; they might work in public or private libraries. While one size never truly fits all in libraries, a common platform and commitment to retraining can be broadly effective across environments.

The network of library professionals, library schools, and cultural heritage organizations that are coming together around the DPLA are also, together, creating a training and retraining system on the national scale. A system such as the DPLA builds upon its network of service hubs – currently up and running in more than one-third of states – to create both a curriculum and training programs to develop skills and capacity within public libraries. Through a series of workshops and online webinars, librarians come together to train one another in the latest new technologies and design techniques. Together, they are building the required skill set in the community while also adding to the availability of materials and code.

This collaborative, distributed approach both to social production and to professional development can work. The idea is to build upon existing relationships and organizations, such as state librarians and their counterparts, wherever possible, avoiding the need to create redundant networks. The outcomes of this networked activity are twofold. First, library students, current librarians, and library volunteers around the country would be trained with new skills, using DPLA-related open materials. Second, these librarians would constantly be creating, through their training, curated exhibits and materials that will be immediately useful to their patrons and to all those who use the DPLA, whether in the United States or elsewhere.

The specific activities in this professional development curriculum include the use of DPLA materials to establish curated exhibits. These locally relevant projects support schools and their libraries in meeting the needs of young patrons. For example, through a DPLA-supported training, a public librarian might establish a customized web environment for the fifth graders in his or her town to understand a scientific phenomenon (such as kinetic energy or the phases of the moon) or a local historical event (such as the California Gold Rush for San Franciscans). The exhibit could be co-developed with local school librarians and teachers to ensure its relevance. If the librarian's work is effective, the resulting metadata could be harvested for broad reuse.

This experiential learning approach can bring librarians in direct contact with some

of the most promising new open-source technologies. For instance, the open-source platform Omeka provides a toolkit for librarians to use as part of the library reinvention process. Omeka is a simple-to-use, inexpensive way to publish digital collections online. The tools that the Omeka team has developed are designed for librarians, archivists, and museum staff who want to curate digital materials into online collections that will entice their patrons. These online collections are networked to one another, as well, such that they can be shared broadly outside the library, archive, or museum where they are created. The work that one librarian does can make the job of the next librarian (seeking to curate a similar or even distantly related exhibit) much easier.[6] The existence of open and free systems such as Omeka is a major reason why this process of reinvention could work today.

The primary advantage of a networked model, supported by a networked community and organized around a common platform, is that it can scale itself sustainably. Think of the extraordinarily quick growth of the Internet and, more recently, the World Wide Web, which is just twenty-five years old. Wikipedia, today one of the world's most frequently visited websites, likewise grew rapidly as a result of this distributed, networked model. The growth of shared resources through a shared platform and social production can sustain libraries and their users for this generation and beyond. New partners – for instance, library and information schools, or related cultural institutions such as museums – could join the network at any time and enable it to expand and grow further.

The constraints and possibilities facing libraries today are important for their own sake: libraries matter a great deal to the proper functioning of our democracy and to our scholarly enterprise. The inquiry into the future of libraries also contributes, though, to our understanding of the nature and importance of public-facing institutions in broader terms. Institutions devoted to the common good, and that have information and knowledge at their core, face similar challenges to one another in a digital era. Libraries are joined by schools and newspapers as institutions under threat of disruption from digital-age competitors; and each also plays an essential role in modern democracies. Profit-seeking competitors to these essential public-facing institutions pose a threat to the extent to which our citizens can inform themselves and participate in effective ways in civic life.

Our risk is that public-spirited services, today provided by libraries, schools, and newspapers and motivated principally by shared interest in the common good, will become the province of profit-driven entities that provide these services less effectively and, perhaps, less ethically. In the field of libraries, there is reason to fear that the runaway success of Google in information retrieval and Amazon in the sale of digital books, music, and movies will draw people away from reliance upon libraries. Even if nominal, this change would cost society in the long run. Libraries and the librarians who work in them serve an essential purpose in a democracy: to provide information and knowledge, free of charge, to people who rely on it as life learners and civic actors. Librarians have no incentive to promote one work over another or one product or service over another; they serve the patrons and their interests solely, and they do so with a fierce commitment to user privacy. These commitments, which are not shared by librarians' commercial counterparts, support the proper functioning of our democracy.

The spheres of education and journalism are closely related to libraries in this respect. Schools and newspapers, like libraries, are under threat from the widespread adoption of digital technologies. Newspapers

already suffer from declining advertising sales and decreased subscription revenues, siphoned off to the web and to search-related advertising. Schools have been less plainly affected to date, but threats from online courses, offered at massive scale, have sent shockwaves through higher education in particular.

The best design choice for leaders in libraries, education, and journalism is the same: resist a world in which the role of public-spirited institutions is dramatically reduced in favor of commercially oriented firms offering slick interfaces. In each field, digital media and networked modes of collaboration can support the core mission of public-facing institutions. In each field, the practice of professionals must change to meet the evolving expectations of those whom we serve and those who pay the bills. In this, the leaders of these institutions must continue to hew to the traditional principles and activities that make their work so effective and important to society: independence, trustworthiness, dependability, and a public-spirited orientation to providing access to knowledge.

The philosophy and methods that have enabled the online world to grow and thrive so quickly can serve librarians, educators, and journalists, too. A commitment to a design sensibility that is oriented toward the needs of users, whether online or in physical spaces, is a crucial starting point. A reorientation toward working not alone, but as networked actors, wherever possible, and using digital platforms in a mode of social production, much as technologists have, is equally important. The nature of the practice in these fields, and of the firm itself, is changing and needs to change, in order for libraries, educators, and journalists to thrive in a digital-plus era.

ENDNOTES

1 Chicago Public Library, "YouMedia," http://www.chipublib.org/youmedia/.

2 Chicago Public Library, "Facts and Figures," http://www.chipublib.org/facts-and-figures/.

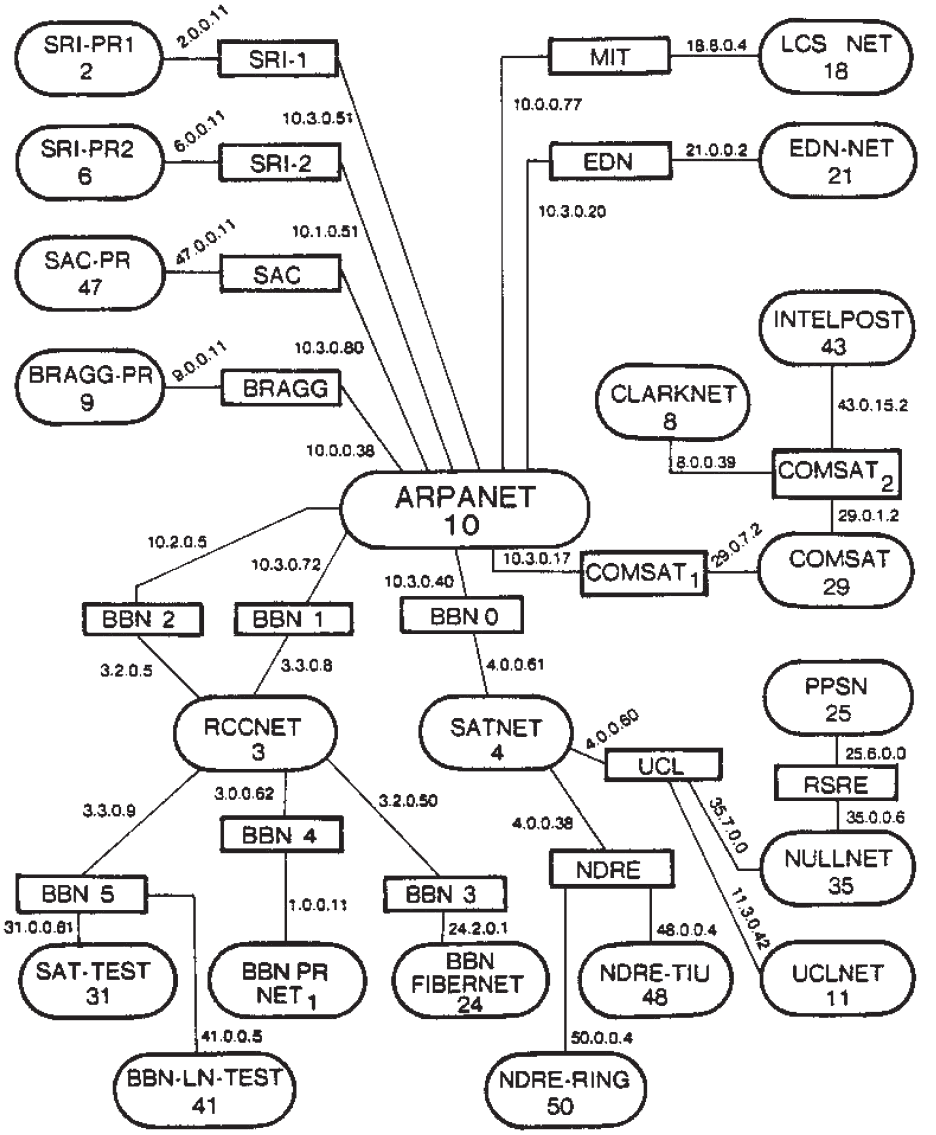
3 Yochai Benkler, *The Wealth of Networks: How Social Production Transforms Markets and Freedom* (New Haven, Conn.: Yale University Press, 2006).

4 Digital Commonwealth (Massachusetts Collections Online), "Little Missouri River in N. Roosevelt Park, North Dakota Badlands," Boston Public Library Tichnor Brothers Postcard Collection, https://www.digitalcommonwealth.org/search/commonwealth:xw42n8060; and Digital Commonwealth (Massachusetts Collections Online), "Ringling Museum, between Sarasota and Bradenton, Florida," Boston Public Library Tichnor Brothers Postcard Collection, https://www.digitalcommonwealth.org/search/commonwealth:41687m16r.

5 Caroline Nolan, "Connected Learning in Practice: The Sidney R. Knafel Map Collection," Tang Institute at Andover, Phillips Academy Andover, http://tanginstitute.andover.edu/2014/04/connected-learning-in-practice-the-sidney-r-knafel-map-collection/.

6 Omeka, http://omeka.org/.

Inside back cover: Reproduction of a slide from an NSA presentation on "Google Cloud Exploitation," an operation of the MUSCULAR surveillance project. The slide is part of the large cache of NSA documents disclosed by Edward Snowden in 2013. The diagram demonstrates the extent to which Internet communications are increasingly managed on private platforms, both mobile and in the cloud. It also illustrates the emergence of new control points in the hands of a small number of private companies and technically capable governments, as well as the associated risks to privacy and security.


SRI-PR1
2
2.0.0.11
SRI-1
SRI-PR2
6
6.0.0.11
SRI-2
SAC-PR
47
47.0.0.11
SAC
BRAGG-PR
9
9.0.0.11
BRAGG
10.3.0.51
10.1.0.51
10.3.0.80
10.0.0.38
ARPANET
10
MIT
18.6.0.4
LCS NET
18
10.0.0.77
EDN
21.0.0.2
EDN-NET
21
10.3.0.20
INTELPOST
43
43.0.15.2
CLARKNET
8
8.0.0.39
COMSAT 2
29.0.1.2
COMSAT
29
29.0.7.2
COMSAT 1
10.3.0.17
10.2.0.5
10.3.0.72
10.3.0.40
BBN 2
BBN 1
BBN 0
3.2.0.5
3.3.0.8
4.0.0.61
RCCNET
3
SATNET
4
4.0.0.60
UCL
PPSN
25
25.6.0.0
RSRE
35.0.0.6
NULLNET
35
35.7.0.0
3.3.0.9
3.0.0.62
3.2.0.50
BBN 4
4.0.0.38
NDRE
BBN 5
BBN 3
1.0.0.11
24.2.0.1
48.0.0.4
31.0.0.61
SAT-TEST
31
BBN PR
NET 1
BBN
FIBERNET
24
NDRE-TIU
48
11.3.0.42
UCLNET
11
41.0.0.5
50.0.0.4
BBN-LN-TEST
41
NDRE-RING
50

coming up in Dædalus:

What's New About the Old?	Matthew S. Santirocco, Caroline Alexander, Roger Bagnall, Shadi Bartsch, Angelos Chaniotis, Greg Crane, Emily Greenwood, Rachel Hadas, Kyle Harper, Brooke Holmes, Phillip Mitsis, Verity Platt, Michael C. J. Putnam, Walter Scheidel & Ian Morris, Peter T. Struck, and Malcolm H. Wiener
On Political Leadership	Archie Brown, Robert Elgie, Nannerl O. Keohane, Eugene Huskey, Alfred Stepan, Barbara Kellerman, Zeynep Tufekci, Alexander Haslam & Stephen D. Reicher, Eric Posner, Michele Swers, and Anthony King
Ethics, Technology & War	Scott D. Sagan, Michael Walzer, Michael Horowitz, David Fidler, Robert Kehler, Jeffrey Lewis & Scott D. Sagan, Lloyd Axworthy & Walter Dorn, Jennifer Leaning, and Benjamin Valentino
The Changing Rules of War	Laura Ford Savarese & John Fabian Witt, Joseph Felter & Jacob N. Shapiro, Allen Weiner, Jennifer Welsh, Tanisha Fazal, Mark Martins & Jacob Bronsther, Leslie Vinjamuri, Keith Krause, Seth Lazar, Janne Nolan & Antonia Chayes, Paul Wise, and Scott D. Sagan

plus Russia: Prospects for Transformative Political Change; Prospects and Limits of Deliberative Democracy; The American Indian: Obstacles and Opportunities &c

U.S. $14; www.amacad.org; @americanacad