JOURNAL ON TELECOMMUNICATIONS & HIGH TECHNOLOGY LAW
is published semi-annually by the
Journal on Telecommunications & High Technology Law,
Campus Box 401, Boulder, CO 80309-0401

ISSN: 1543-8899

Copyright © 2004 by the
Journal on Telecommunications & High Technology Law
an association of students sponsored by the
University of Colorado School of Law and the
Silicon Flatirons Telecommunications Program.

POSTMASTER: Please send address changes to JTHTL,
Campus Box 401, Boulder, CO 80309-0401

Subscriptions

Volume subscriptions are available for $ 45.00. City of Boulder subscribers please add $3.45 sales tax. Boulder County subscribers outside the City of Boulder please add $1.91 sales tax. Metro Denver subscribers outside of Boulder County please add $1.67 sales tax. Colorado subscribers outside of Metro Denver please add $1.31 sales tax.

Inquiries concerning ongoing subscriptions or obtaining an indvidual issue should be directed to the attention of JTHTL Managing Editor at JTHTL@colorado.edu or by writing JTHTL Managing Editor, Campus Box 401, Boulder, CO 80309-0401.

Back issues in complete sets, volumes, or single issues may be obtained from: William S. Hein & Co., Inc., 1285 Main Street, Buffalo, NY 14209.

Manuscripts

JTHTL invites the submission of unsolicited manuscripts. Please send softcopy manuscripts to the attention of JTHTL Articles Editors at JTHTL@colorado.edu in Word or PDF formats or through ExpressO at http://law.bepress.com/expresso. Hardcopy submissions may be sent to JTHTL Articles Editors, Campus Box 401, Boulder, CO 80309-0401. Unfortunately, JTHTL cannot return manuscripts. JTHTL uses THE BLUEBOOK: A UNIFORM SYSTEM OF CITATION (17th ed. 2000) for citation format and THE CHICAGO MANUAL OF STYLE (15th ed. 2003) for a style guide.

Cite as: 3 J. ON TELECOMM. & HIGH TECH. L. __ (2004).

J. ON TELECOMM. & HIGH TECH. L.

JOURNAL ON TELECOMMUNICATIONS & HIGH TECHNOLOGY LAW

Volume 3 *Fall 2004*

BOARD OF EDITORS

EDITOR IN CHIEF
Eric D. Gunning

MANAGING EDITOR
Scott A. Goodwin

PRODUCTION EDITOR
Emily D. Lauck

ARTICLES EDITORS
Joel Dion
Cory Jackson
Andrew D. Johnson

CASENOTE & COMMENT EDITORS
Kley Achterhof
Chelsea May
Jennifer L. Owens
Andrew Teske

EXECUTIVE EDITOR
Karl A. Dierenbach

ASSISTANT PRODUCTION EDITOR
Paul Dunlap

ASSOCIATE EDITORS
Thomas Blumstrom
Peter Chen
Jason Mashek
Jonathon Blum
Jamey Green
Tim Newlin
Xiaolu Zhang

MEMBERS
Molly Ferrer
Lisa Neal-Graves
Andrew Hogle
Andrew LaFontaine
Jennifer Loyd
Alexander Ross
Paul Frinak
Berkeley Harris
Todd Hoy
Zachary Lange
Alison Minea
Rita Sanzgiri
Cynthia Sweet
Joshua Graae
Patricia Ho
Heather Kenney
Travis Litman
Christopher Myers
Margot Summers

FACULTY ADVISOR
Philip J. Weiser, *Associate Professor of Law*
Executive Director of the Silicon Flatirons Telecommunications Program

J. ON TELECOMM. & HIGH TECH. L.

THE UNIVERSITY OF COLORADO SCHOOL OF LAW

FACULTY, 2004-05

BARBARA A. BINTLIFF, *Nicholas Rosenbaum Professor of Law and Law Library Director.* B.A., Central Washington State College; J.D., M.L.L., University of Washington.

HAROLD H. BRUFF*, Charles Inglis Thomson Professor of Law.* B.A., Williams College; J.D., Harvard University.

CLIFFORD J. CALHOUN, *Professor Emeritus.* A.B., LL.B., Harvard University.

EMILY M. CALHOUN, *Professor of Law.* B.A., M.A., Texas Tech University; J.D., University of Texas.

PAUL F. CAMPOS, *Professor of Law.* A.B., M.A., J.D., University of Michigan.

HOMER H. CLARK, JR., *Professor Emeritus.* A.B., LL.D., Amherst College; LL.B., LL.M., Harvard University.

RICHARD B. COLLINS, *Professor of Law and Director of the Byron R. White Center for the Study of American Constitutional Law.* B.A., Yale College; LL.B., Harvard University.

JAMES N. CORBRIDGE, JR., *Professor Emeritus.* A.B., Brown University; LL.B., Yale University.

NESTOR DAVIDSON, *Associate Professor of Law.* A.B., Harvard University; J.D., Columbia University.

RICHARD DELGADO, *Jean N. Lindsley Professor of Law.* A.B., University of Washington; J.D., University of California, Berkeley.

ALLISON HARTWELL EID, *Associate Professor of Law.* A.B., Stanford University; J.D., University of Chicago.

TED J. FIFLIS, *Professor of Law.* B.S., Northwestern University; LL.B., Harvard University.

WAYNE M. GAZUR, *Professor of Law.* B.S., University of Wyoming; J.D., University of Colorado; LL.M., University of Denver.

DAVID H. GETCHES, *Dean and Raphael J. Moses Professor of Natural Resources Law.* A.B., Occidental College; J.D., University of Southern California.

LAKSHMAN GURUSWAMY, *Professor of Law.* LL.B., Sri Lanka; Ph.D., University of Durham, U.K.

MELISSA HART, *Associate Professor of Law.* B.A., Harvard-Radcliffe College; J.D., Harvard University.

DAVID S. HILL, *Professor of Law.* B.S., J.D., University of Nebraska.

CLARE HUNTINTON, *Associate Professor of Law.* B.A., Oberlin College; J.D., Columbia University.

J. DENNIS HYNES, *Nicholas Rosenbaum Professor Emeritus.* B.A., LL.B., University of Colorado.

HOWARD C. KLEMME, *Professor Emeritus*. B.A., LL.B., University of Colorado; LL.M., Yale University.

SARAH A. KRAKOFF, *Associate Professor of Law*. B.A., Yale University; LL.B, University of California, Berkeley.

MARK J. LOEWENSTEIN, *Associate Dean for Research and Professor of Law*. A.B., J.D., University of Illinois.

DAYNA BOWEN MATTHEW, *Associate Professor of Law,* A.B., Harvard; J.D., University of Virginia.

CHRISTOPHER B. MUELLER, *Henry S. Lindsley Professor of Procedure and Advocacy*. A.B., Haverford College; J.D., University of California, Berkeley.

ROBERT F. NAGEL, *Ira C. Rothgerber, Jr. Professor of Constitutional Law*. B.A., Swarthmore College; J.D., Yale University.

DALE OESTERLE, *Monfort Professor of Commercial Law and Director of the Entrepreneurial Law Center*. B.A., M.P.P., J.D., University of Michigan.

SCOTT R. PEPPET, *Associate Professor of Law*. B.A., Cornell University; J.D., Harvard University.

COURTLAND H. PETERSON, *Nicholas Doman Professor of International Law Emeritus*. B.A., LL.B., University of Colorado; M. Comp. L., University of Chicago; Dr. Jur., University of Freiburg (Germany).

WILLIAM T. PIZZI, *Professor of Law*. A.B., Holy Cross College; M.A., University of Massachusetts; J.D., Harvard University.

CAROLYN B. RAMSEY, *Associate Professor of Law*. B.A., University of California, Irvine; A.M., Stanford University; J.D., Stanford University.

KEVIN R. REITZ, *Professor of Law*. B.A., Dartmouth College; J.D., University of Pennsylvania.

WILLIAM E. RENTFRO, *Professor Emeritus*. B.A., University of Colorado; Th.M., LL.B., University of Denver.

PIERRE J. SCHLAG, *Byron White Professor of Law*. B.A., Yale University; J.D., University of California, Los Angeles.

AMY J. SCHMITZ, *Associate Professor of Law*. B.A., Drake University; J.D., University of Minnesota.

DON W. SEARS, *Professor Emeritus*. B.S., J.D., Ohio State University.

PETER N. SIMON, *Associate Professor Emeritus*. B.S., M.D., University of Wisconsin; J.D., University of California, Berkeley.

NORTON L. STEUBEN, *Nicholas Rosenbaum Professor of Law Emeritus*. A.B., J.D., University of Michigan.

ARTHUR H. TRAVERS, JR., *Professor Emeritus*. B.A., Grinnell College; LL.B., Harvard University.

MICHAEL J. WAGGONER, *Associate Dean for Academic Affairs and Associate Professor of Law*. A.B., Stanford University; LL.B., Harvard University.

PHILIP J. WEISER, *Associate Professor of Law and Executive Director of the Silicon Flatirons Telecommunications Program*. B.A., Swarthmore College; J.D., New York University.

MARIANNE WESSON, *Professor of Law and Wolf-Nichol Fellow*. A.B., Vassar College; J.D., University of Texas.

AHMED A. WHITE, *Associate Professor of Law.* B.A., Southern University and A & M College; J.D., Yale University.

CHARLES F. WILKINSON, *University's Distinguished Professor and Moses Lasky Professor of Law.* B.A., Denison University; LL.B., Stanford University.

SIENHO YEE, *Associate Professor of Law.* Peking University, B.A., Brandeis University; J.D., Columbia University; University of Oxford.

Research and Clinical Faculty

NORMAN F. AARONSON, *Clinical Professor*, *Legal Aid and Defender Program.* A.B., Brandeis University; J.D., Boston University.

ROBERT J. DIETER, *Clinical Professor*, *Legal Aid and Defender Program.* B.A., Yale University; J.D., University of Denver.

H. PATRICK FURMAN, *Clinical Professor, Legal Aid and Defender Program, and Director of Clinical Programs.* B.A., J.D., University of Colorado.

JULIET C. GILBERT, *Clinical Professor, Legal Aid and Defender Program.* B.A., Valparaiso University; J.D., University of Denver.

JILL E. TOMPKINS, *Instructor and Director of the Indian Law Clinic.* B.A., The King's College; J.D., University of Maine.

Law Library Faculty

BARBARA A. BINTLIFF, *Nicholas Rosenbaum Professor of Law and Law Library Director.* B.A., Central Washington State College; J.D., M.L.L., University of Washington.

GEORGIA K. BRISCOE, *Associate Director and Head of Technical Services.* B.S., Washington State University; M.A., University of San Diego; M.L.S., University of Michigan.

DONALD L. FORD, *Reference Librarian.* B.A., American University School of International Service; J.D., University of Virginia; M.L.I.S., University of Pittsburgh School of Information Sciences.

DRUET CAMERON KLUGH, *Reference Librarian.* B.A., J.D., University of Iowa.

KAREN SELDEN, *Catalog Librarian.* B.S., Pennsylvania State University; M.L.S., Simmons College.

YUMIN JIANG, *Technical Services Librarian.* M.S., University of Illinois, Urbana-Champaign; M.A. University of Wisconsin.

RUSSELL SWEET, *Head of Public Services.* B.A, University of California, Riverside; MAR, Yale University; J.D., University of Washington; M.L., University of Washington.

JANE E. THOMPSON, *Head of Faculty Services.* B.A., University of Missouri; M.A., J.D., University of Denver.

Legal Writing and Appellate Advocacy Faculty

LOUISA HEINY, *Legal Writing Instructor.* B.A., J.D., University of Colorado.

NATALIE MACK, *Legal Writing Instructor.* B.S., University of South Carolina; J.D., University of Colorado.

GABRIELLE M. STAFFORD, *Legal Writing Professor.* B.A., University of Pennsylvania; J.D., Boston University.

TODD M. STAFFORD, *Legal Writing Professor.* B.A., Southern Methodist University; J.D., Duke University.

Research Associates

DOUGLAS S. KENNEY, *Research Associate, Natural Resources Law Center.* B.A., University of Colorado; M.S., University of Michigan School of Natural Resources and Environment; Ph.D., Cornell University.

KATHRYN M. MUTZ, *Research Associate, Natural Resources Law Center.* B.A., University of Chicago; M.S., Utah State University; J.D., University of Colorado.

JEAN STEFANCIC*, Senior Research Associate.* B.A., Maryville College; M.L.S., Simmons College; M.A., University of San Francisco.

Adjunct, Adjoint and Visiting Faculty

GARRY R. APPEL, *Attorney at Law, Appel & Lucas, P.C., Denver, Colorado.* B.A., J.D., University of Colorado.

GEORGE BRAUCHLER, *Deputy District Attorney, First Judicial District, Golden, Colorado.* B.A., J.D., University of Colorado.

SHARON CAULFIELD, *Attorney at Law, Caplan & Earnest, LLC, Boulder, Colorado.* B.A., J.D., University of Colorado.

CHRISTIE COATES, *Attorney at Law, Boulder, Colorado.* B.A., Houston Baptist University; M.Ed., University of Houston; J.D., University of Colorado.

SEAN CONNELLY, *Partner, Hoffman, Reilly, Pozner & Williamson, Denver, Colorado.* A.B., Fairfield University; J.D., Catholic University Law School.

STEVEN CLYMER, *Attorney at Law, ACCORD Dispute Resolution Services, Boulder, Colorado.* A.B., St. Louis University; J.D., Case Western Reserve University.

WILEY DANIEL, *Judge, United States District Court for the District of Colorado.* B.A., J.D., Howard University.

DANIEL DEASY, *Attorney at Law, George Browning & Associates, Westminster, Colorado.* B.A, J.D., University of Colorado.

ROGER FLYNN, *Executive Director, Western Mining Action Project, Boulder, Colorado.* B.S., Lehigh University; J.D., University of Colorado.

JOHN A. FRANCIS, *Partner, Davis, Graham, & Stubbs, Denver, Colorado.* B.A., University of Colorado; J.D., University of Michigan.

EDWARD J. GAC, *Associate Professor of Taxation and Business Law, College of Business, University of Colorado, Boulder.* A.A., Wright College; B.A., Western Illinois University; J.D., University of Illinois.

CRAIG C. GARBY, *Associate, Gibson, Dunn & Crutcher, LLP, Denver, Colorado.* B.A., University of Colorado; Graduate Research, Waseda University, Tokyo, Japan; M.P.A., Cornell University; J.D., Stanford University.

JASON D. HAISLMAIER, *Associate, Holme Roberts & Owen LLP, Boulder, Colorado.* B.S., Northwestern University; J.D., Franklin Pierce Law Center.

ANDREW HARTMAN, *Attorney at Law, Cooley Godward, LLP, Broomfield, Colorado.* A.B., University of Michigan; J.D., Georgetown University.

BETTY JACKSON, *Professor of Accounting, School of Business, University of Colorado, Boulder.* BBA, Southern Methodist University; M.P.A., Ph.D., University of Texas, Austin.

THOMAS D. LUSTIG, *Senior Staff Attorney, National Wildlife Federation, Boulder, Colorado.* A.B., Washington University; M.S., University of Michigan; J.D., University of Colorado; Ph.D., Massachusetts Institute of Technology.

JACK MILLS, *Attorney at Law, A.J. Mills, P.C., Boulder, Colorado.* BBA, LL.B., University of Oklahoma.

VIVA R. MOFFAT, *Attorney at Law, Law Offices of David Mastbaum, Boulder, Colorado.* A.B., Stanford University; M.A., J.D., University of Virginia.

ANN MORGAN, Adjoint Professor, University of Colorado, Boulder, Colorado. B.S., University of California, Berkeley; M.B.A., Golden Gate University.

RUTH ORATZ, *Genetic Counselor, Rocky Mountain Cancer Center, Denver, Colorado.* A.B., Harvard University; M.D., Albert Einstein College of Medicine

CHRISTOPHER D. OZEROFF, *Partner, Hogan & Hartson LLP, Boulder, Colorado.* B.A., Stanford University; J.D., University of Chicago.

DOROTHY RAYMOND, *Senior Vice President and General Counsel, CableLabs, Denver, Colorado.* B.A., University of Denver; J.D., University of Colorado.

THE HONORABLE NANCY E. RICE, *Justice, Colorado Supreme Court, Denver, Colorado.* B.A., Tufts University; J.D., University of Utah.

THE HONORABLE EDWARD J. RICHARDSON, *State of Florida Circuit Court Judge, Retired.* A.S., Brevard Community College; B.S., University of Florida; J.D., Florida State University.

Patrick Ryan, *Attorney at Law, P.S.R. Lawfirm, Denver, Colorado.* B.A., M.B.A., Monterey Institute of International Studies; J.D., University of Texas at Austin; M.B.L., Universität St. Gallen, Switzerland; Ph.D. Katholieke Universiteit Leuven, Belgium.

Micael Saul, *Attorney, National Wildlife Federation, Boulder, Colorado.* B.A., J.D., Yale University.

Stuart W. Stuller, *Attorney at Law, Caplan & Earnest, Boulder, Colorado.* B.A., University of Wisconsin; J.D., University of Colorado.

Karen Taylor, *Deputy Public Defender, Colorado State Public Defender Office, Denver, Colorado.* B.A., Missouri Southern State College; J.D., Northwetern University.

Nathaniel Trelease, *President, WebCredenza, Inc., Denver, Colorado*. B.S., University of Wyoming; J.D., University of Wyoming; LL.M, University of Denver.

Deanna Westfall, *Attorney at Law, Bennington Johnson Biermann & Craigmile LLC, Denver, Colorado*. B.A., Washington College, St. Louis; J.D., University of Colorado.

FROM THE EDITOR

Entering our third year of publication, there is cause for excitement at the *Journal on Telecommunications and High Technology Law* (JTHTL). In addition to making the long awaited move to two issues per year, we are continuing JTHTL's valuable contribution to the ongoing debate over the future of communications policy. Working in close coordination with the Silicon Flatirons Telecommunications Program, JTHTL's mission is to bring a refreshing and innovative perspective to the array of issues arising because of the "Great Digital Broadband Migration."[1]

This first issue of volume three continues JTHTL's mission by furthering the debate on a number of intriguing issues confronting regulators, academics, and industry participants. This issue begins with a compelling plea from FCC Chairman Michael K. Powell to the industry urging the voluntary adoption of a set of "Internet Freedoms," principles that would guarantee customer access to Internet content free from any arbitrary restrictions imposed by broadband providers. Following Chairman Powell's speech, Professors Christopher Yoo and Tim Wu cogently debate the pros and cons of a government-mandated network neutrality regime. These two articles provide a glimpse into the well-articulated arguments that characterized last spring's Silicon Flatirons Telecommunications Program Conference.

Next, Professor Howard Shelanski examines the competitive landscape and role of antitrust policy in the deployment of the third broadband "pipe" to the home: 3G Wireless technologies. The following two articles in this issue provide thoughtful analysis to a crucial, albeit little discussed, area of telecommunications and Internet law. Scott Marcus of the FCC's Office of Strategy Planning and Policy Analysis addresses the appropriate level of government intervention required in upgrading the Internet's infrastructure, while Professor Peter Swire provides a theoretical perspective on the propriety of disclosing vulnerabilities in order to increase security on the Internet. Finally, Volume 3, Issue 1 concludes with the 2003 Silicon Flatirons Student Writing Contest Winner, Joe Linhoff, who discusses the implications of

[1] See Michael K. Powell, *Preserving Internet Freedom: Guiding Principles for the Industry*, 3 J. ON TELECOMM. & HIGH TECH. L. 5, 5 & n.1 (2004).

the Digital Millennium Copyright Act on reverse engineering in the video game industry.

Bringing these debates to publication requires countless hours of work and dedication from the entire journal staff. Although the entire Board deserves my utmost gratitude, there are a few individuals who I would like to thank personally for making my life easier and production of this issue possible. I would especially like to thank Emily Lauck, my Production Editor, for all of her hard work in organizing cite checks, helping to produce the Journal, and manning the helm while Scott Goodwin, my Managing Editor, and I were in Washington, D.C. this summer. As for Scott, I would like to thank him for his ability to take seemingly insurmountable problems and place them in a context that made overcoming them effortless. I would also like to thank Karl Dierenbach and my Articles Editors (Cory Jackson, Andrew Johnson, and Joel Dion) for working so diligently without a complaint even when my requests may have been unreasonable.

There are others, outside the journal staff, who deserve recognition. First, I am indebted to the eight authors in this issue for understanding that with a relatively new journal, there will be inevitable growing pains. Their understanding and flexibility made the production process much easier. Second, the current JTHTL Board are mere caretakers holding the fate of this journal in trust for those who laid the groundwork before us and those who will carry the torch after we leave. For the past members of the journal, thank you for making a Volume 3 possible and to those that follow, I wish you success in continuing what we began.

Finally, and most importantly, I must express my sincere gratitude for everything that Professor Phil Weiser has done for both this journal and for its members. A few years ago, Phil Weiser breathed life into the University of Colorado School of Law by establishing both the Silicon Flatirons Telecommunications Program as well as JTHTL. As a product of his inspiration and dedication, Phil Weiser deserves all the credit for this journal's continued success. I can confidently speak for every member of the journal in saying that we are each indebted to Professor Weiser for his guidance and friendship. Personally, I can never repay him for all the doors he has opened and advice he has given.

With that being said, I am proud to publish Volume 3, Issue 1 of the *Journal on Telecommunications and High Technology Law* and am confident that this Journal will continue to bring a refreshing approach to the intellectual debates that make the telecommunications industry the most interesting, stimulating, and dynamic area of the law today.

Eric D. Gunning
Editor-In-Chief

JOURNAL ON TELECOMMUNICATIONS & HIGH TECHNOLOGY LAW

Volume 3 *Fall 2004*

CONTENTS

THE DIGITAL BROADBAND MIGRATION:
TOWARD A REGULATORY REGIME FOR THE INTERNET AGE

A symposium co-sponsored by the *Journal on Telecommunications and High Technology Law* and the Silicon Flatirons Telecommunications Program

INTRODUCTION: A REGULATORY REGIME FOR THE INTERNET AGE

PHILIP J. WEISER[*]

In November of 2000, then-Commissioner Michael K. Powell spoke at the University of Colorado School of Law to discuss the implications of the "digital broadband migration."[1] The pace of this migration continues to accelerate. Indeed, it seems quite likely that we will look back at the years between 2000-2010 as consumed—at least in telecommunications policy circles—by questions related to how to address the broadband Internet. At present, however, we are only glimpsing the beginnings of broadband deployment, the development of security for an evolving infrastructure, and the relationship between broadband providers and complementary applications (such as Voice over Internet Protocol (VoIP)) that ride on top of them.

The set of papers published in this issue of the *Journal on Telecommunications and High Technology Law* (JTHTL) reflects the effort by the Silicon Flatirons Telecommunications Program to raise the level of the debate on cutting edge technology policy questions. With respect to the questions raised by the broadband Internet, the JTHTL is off to a promising start. Notably, its first issue has spurred an important—and ongoing—debate about the virtues of a layered model for telecommunications policy.[2] This issue continues that tradition by addressing the challenging questions regarding whether regulation

[*] Associate Professor of Law and Telecommunications and Executive Director of the Silicon Flatirons Telecommunications Program, University of Colorado.

[1] He later published those remarks as delivered to the Progress and Freedom Foundation. *See* Michael K. Powell, *Preserving Internet Freedom: Guiding Principles For The Industry*, 3 J. ON TELECOMM. & HIGH TECH. L. 5, 5 n.1 (2004).

[2] *See* Philip J. Weiser, *Law and Information Platforms*, 1 J. ON TELECOMM. & HIGH TECH. L. 1, 12 n.51 (2002); Kevin Werbach, *A Layered Model for Internet Policy*, 1 J. ON TELECOMM. & HIGH TECH. L. 37, 38 (2002); Douglas C. Sicker & Joshua L. Mindel, *Refinements of a Layered Model for Telecommunications Policy*, 1 J. ON TELECOMM. & HIGH TECH. L. 69, 71 (2002); John T. Nakahata, *Regulating Information Platforms: The Challenges of Rewriting Communications Regulation from the Bottom Up*, 1 J. ON TELECOMM. & HIGH TECH. L. 95, 98 (2002); *see also* Richard S. Whitt, *A Horizontal Leap Forward: Formulating a New Communications Public Policy Framework Based on the Network Layers Model*, 56 FED. COMM. L.J 587 (2004) (citing heavily to papers published in Volume 1 of the JOURNAL ON TELECOMMUNICATIONS AND HIGH TECHNOLOGY LAW).

should seek to preserve the Internet's open architecture and how policymakers should approach issues related to Internet security.

The four papers in this issue addressing broadband policy grapple with some of the most difficult and most important questions related to the digital broadband migration. In many respects, the fundamental promise of the broadband era is that all sorts of applications—whether VoIP, video on demand, electronic commerce, or those not yet invented—can be provided over broadband connections. Ideally, the rise of broadband Internet platforms will eviscerate the legacy distinctions between different platforms (wired telephone, wireless, cable, etc.) and facilitate entry by innovative application providers. But this vision is by no means assured, as incumbents might—in an attempt to protect their legacy business models—seek to use regulation or exclusionary conduct to limit entry. As one observer remarked, incumbent broadband providers might respond to the threat presented by Vonage, a leading VoIP provider, by using the "dodgy competitive tactic" of "slow[ing] down Vonage's service" as well as "give network precedence to their own revenue-generating services."[3]

Policing anticompetitive conduct in the broadband Internet age will present regulators with the challenge of reorienting their analytical frameworks for a new technological and economic environment. In particular, as Joseph Farrell and I have explained, the economics of vertical integration in this environment are far more complex than many policymakers appreciate.[4] Recognizing this complexity, Chairman Powell announces—in this issue—an "Internet Freedom" policy that puts broadband providers on notice that any departures from non-discrimination norms (i.e., favoring their vertically integrated affiliates) will be frowned-upon. This "jawboning" and enlightened guidance to the industry is, however, likely only to postpone the day when the Federal Communications Commission (FCC) is forced to evaluate the competitive consequences of discrimination that arises from vertical integration.[5]

In this issue, Christopher Yoo and Tim Wu evaluate the arguments, albeit from different perspectives, that bear on the competitive effects of

[3] Daniel Klein, *Why Vonage Is Just A Fad*, ZDNET (May 19, 2004), *available at* http://techupdate.zdnet.com/techupdate/stories/main/Why_Vonage_Just_Fad.html?tag=tu.arch.link.

[4] Joseph Farrell & Philip J. Weiser, *Modularity, Vertical Integration, and Open Access Policies: Towards a Convergence of Antitrust and Regulation in the Internet Age*, 17 HARV. J.L. TECH. 85, 105-19 (2003).

[5] "Jawboning" refers to statements by policymakers that threaten possible action, as opposed to announcing actual action.

discrimination between applications riding on broadband networks.[6] Indeed, their point-counterpoint effectively illustrates how the Internet Freedom debate often turns on epistemological grounds.[7] By that, I mean that one's basic premise of "How do I know what I think I know?" will often dictate one's approach to the Internet freedom issue. Thus, for those believing that the Internet's modularity and historic openness is critically important to facilitating entry and innovation, the need for FCC action is obvious. By contrast, for those believing that vertical integration generally facilitates valuable efficiencies and spurs new investment, the need for regulatory restraint is equally obvious. For the rest of us (i.e., those uncertain of the primacy of either asserted position), it is far from obvious how to confront this policy challenge.

As a general matter, I resolve the challenge of how to address the competitive effects of vertical integration (and any associated discrimination towards certain application providers) by using an antitrust model of regulation. In this respect, I share Howard Shelanski's endorsement of sector-specific regulation when addressing "the oversight of interconnection and its associated pricing issues."[8] In particular, I endorse an antitrust-like model of regulation as a means of sorting the wheat from the chaff—in terms of identifying exclusionary discrimination—and addressing the questions that the FCC will face when and if it is forced to take a formal stand on the issue (i.e., if the jawboning strategy is not a viable long term approach).[9] In that regard, I must note that there are other possible approaches, such as admonishing broadband providers to adopt clear policies towards application providers and to enforce those policies at the FCC in a manner similar to how the Federal Trade Commission enforces Internet privacy policies.[10] Indeed, both because of the complexity of this issue and the different permutations of possible regulatory responses, Internet Freedom issues are likely to be debated for some time. And regardless of how that debate ends, I am confident that the articles in this issue will elevate that discussion and help point the way towards an effective solution.

[6] *See* Christopher S. Yoo, *Would Mandating Broadband Network Neutrality Help or Hurt Competition? A Comment on the End-to-End Debate*, 3 J. ON TELECOMM. & HIGH TECH. L. 23 (2004); Tim Wu, *The Broadband Debate, a User's Guide*, 3 J. ON TELECOMM. & HIGH TECH. L. 69 (2004).

[7] "Epistemological" refers to the branch of philosophy that studies "the nature of knowledge."

[8] Howard A. Shelanski, *Competition Policy for Mobile Broadband Networks*, 3 J. ON TELECOMM. & HIGH TECH. L. 97, 118 (2004).

[9] *See* Philip J. Weiser, *Toward A Next Generation Regulatory Regime*, 35 LOY. L. REV. 41 (2003).

[10] *See* Steven Hetcher, *The FTC As An Internet Privacy Norm Entrepeneur*, 53 VAND. L. REV. 2041 (2000).

Like the issues related to broadband policy, the questions swirling around security policy beg for thoughtful analysis. To date, legal scholars have largely avoided this intimidating set of issues. Thankfully, Peter Swire, one of the leading scholars in this area, is an exception to the rule, as evidenced by his thoughtful analysis of the disclosure of security vulnerabilities.[11] Similarly, Scott Marcus, whose technical training shows through in his article, provides an important analysis discussing how the development of the Internet can address security concerns.[12] These two perspectives, however, reflect only the very beginnings of the debate over the security policy, which is now roughly at the stage that the broadband policy debate was in 2000. In future offerings, the JTHTL will strive to publish more scholarship in this area and help advance what is almost certain to become an increasingly important area of technology policy.

[11] *See* Peter P. Swire, *A Model for When Disclosure Helps Security: What is Different About Computer and Network Security?*, 3 J. ON TELECOMM. & HIGH TECH. L. 163 (2004).

[12] *See* J. Scott Marcus, *Evolving Core Capabilities of the Internet*, 3 J. ON TELECOMM. & HIGH TECH. L. 121 (2004).

PRESERVING INTERNET FREEDOM: GUIDING PRINCIPLES FOR THE INDUSTRY

MICHAEL K. POWELL[*]

I. THE VISION FOR THE BROADBAND INTERNET

I want to thank Professor Phil Weiser and the University of Colorado School of Law for letting me speak here today on the "Digital Migration," a term I introduced here at the Silicon Flatirons Conference four years ago, to describe our movement from a slow, conventional, analog world, to a digital world that promises so many incredible opportunities for faster, more reliable, and higher-quality communications.[1] The move to this digital world is a radical transformation and its benefits will be felt by each and every American.

Those of you who follow the Federal Communications Commission (Commission) closely should be very familiar with the agency's vision for

[*] This essay was adapted from a speech and question and answer session delivered by FCC Chairman Michael K. Powell at the Symposium on "The Digital Broadband Migration: Toward a Regulatory Regime for the Internet Age" held at the University of Colorado School of Law on February 8, 2004.

[1] A few days after addressing the Silicon Flatirons Conference at the University of Colorado School of Law, then Commissioner Powell gave the speech to a convention at the Progress and Freedom Foundation. FCC Commissioner Michael K. Powell, The Great Digital Broadband Migration, Address before the Progress & Freedom Foundation (Dec. 8, 2000) (transcript available at http://www.fcc.gov/Speeches/Powell/2000/spmkp003.html).

the high-speed, broadband Internet. Our national broadband policy seeks to promote investment in diverse, faster, and more sophisticated Internet and related technologies.[2] This, in turn, will foster economic growth, innovation, and empower American consumers to make more choices in how they live, work, and play.

Indeed, a recent Pew Internet Study suggests that consumers are already beginning to take advantage of the new opportunities provided by high-speed connections in their homes.[3] According to the report, those with broadband generally do more online than those with dial-up connections.[4] This includes peer-to-peer file sharing, enhanced instant messaging, streaming video, and using virtual private networks.

The next generation of broadband will make both new applications possible and established applications more compelling. But we will not get there through wishful thinking. Everyone involved in the broadband Internet – end-users, network providers, content producers, applications developers, and policymakers – must continue to be missionaries in driving infrastructure and applications deployment if our nation hopes to stand among the leaders of the Information Age.

To date, experiments in dial-up access have given Americans a growing number of ways to communicate, gather information, and be entertained. High-speed Internet accelerates that trend. These expanded choices, in turn, result in lower prices and higher value. In addition, the almost infinite flexibility of the Internet Protocol gives users the tools to tailor these valuable innovations to their own individual needs – to make them their own.

All this activity is precisely what Congress had in mind when it directed the Commission to "encourage the deployment [of broadband] on a reasonable and timely basis . . . ."[5] The Act also mandates we take "action to accelerate deployment."[6] We have and we will.

That is why the Commission has pushed so hard to create incentives and tools to encourage companies to bring consumers additional high-speed Internet technologies. We have taken steps to promote investment in traditional platforms, like cable modems and

[2] For an outline of the FCC's policies and objectives regarding broadband, see FCC, BROADBAND, *at* http://www.fcc.gov/broadband/ (last reviewed/updated March 13, 2003).

[3] MARY MADDEN, AMERICA'S ONLINE PURSUITS: THE CHANGING PICTURE OF WHO'S ONLINE AND WHAT THEY DO (Pew Internet & Am. Life Project, Report, Dec. 22, 2003), *available at* http://www.pewInternet.org/reports/pdfs/PIP_Online_Pursuits_Final.PDF.

[4] *Id.* at 5.

[5] Telecommunications Act of 1996, Pub. L. 104-104, 110 Stat. 56, § 706(a) (codified at 47 U.S.C. § 157(a)).

[6] *Id.*

DSL.[7] We are particularly proud, however, that we are leading the charge for new, emerging broadband platforms, such as broadband over power lines, WiFi and WiMax, Ultra-wideband, satellite, and the list goes on.

Of course, a real bright spot has been the hot spot. By making licensed and unlicensed wireless spectrum available for broadband uses, we have seen an explosion of wireless access points and have witnessed blossoming wireless technologies that allow powerful, untethered Internet access around the country. As we look forward, our goal is to continue to champion and facilitate the higher-speed, more capable platforms and applications of tomorrow.

These efforts to promote investment and competition among networks follow from a simple truth: no amount of regulation, or wishful thinking, will bring consumers the benefits of high-speed Internet if the networks are not in place to serve them. We have an historic opportunity to bring multiple pipes to consumers, and, thereby, take a big bite out of the "last mile" problems that have plagued competition for a century and invited, almost necessarily, heavy monopoly regulation.

II. ACHIEVING THE VISION: POWER TO THE PEOPLE

Promoting competition among high-speed Internet platforms, however, is only half of the task. We have to ensure that these technologies' various capabilities are not used in a way that could stunt the growth of the economy, innovation, and consumer empowerment. Thus, we must expand our focus beyond broadband networks – the so-called "physical layer" of the Internet's layered architecture.

Again, broadband networks are impressive generators of economic growth, innovation, and empowerment. But generators do not work unless they have *fuel to burn*. Broadband networks are fueled by consumers' hunger for an ever-expanding array of high-value content, applications, and personal devices that can run over these networks. Easy access to content and technology is bringing more power to people.

Personal computing devices are at the leading edge of this revolution in consumer empowerment. These devices exploit the rapid innovation in silicon, software, and storage, and combine it with speedy Internet connections. This potent combination is putting in the hands

[7] *See* Appropriate Framework for Broadband Access to the Internet over Wireline Facilities, *Notice of Proposed Rulemaking*, 17 FCC Rcd. 3019, 3023, at ¶ 5 (2002) (stating that "broadband services should exist in a minimal regulatory environment that promotes investment and innovation in a competitive market"); Inquiry Concerning High-Speed Access to the Internet Over Cable & Other Facilities, *Declaratory Ruling & Notice of Proposed Rulemaking*, 17 FCC Rcd. 4798, 4802, at ¶ 5 (2002), *aff'd in part, vacated in part by* Brand X Internet Servs. v. FCC, 345 F.3d 1120 (9th Cir. 2003) (same).

of Americans the same computing power that once was reserved for CalTech, the military, or the phone company.

You have no doubt heard the litany of electronic devices that can offer consumers more options and more personalization using the Internet: just open your paper and look at the advertisements on Sunday morning. Music players like the iPod; personal video recorders like TiVo; boxes for Internet voice from service providers like Vonage; online game systems like Xbox and GameCube; smart phones; and WiFi that allows you to surf the Internet from your local coffee shop or your back porch are the common statements of our culture today.

But the possibilities for consumer empowerment extend beyond just your gadgets. Those possibilities arise from the Internet's open architecture, which allows consumers to freely interact with anyone around the globe. Musicians and writers, who never could have landed a contract with a major record label or a publisher, can now find an audience for their work. A small town radio station serving a dwindling audience can suddenly reach a market that has moved to the big city. Take eBay, for example: gone are the days when each of us had only a small group of potential buyers for what we thought was junk in our garages. Somewhere, in the next state or maybe the next continent, there are people who may very well want to buy that "junk" and pay us more than we ever dreamed for it. The open Internet has opened markets beyond the traditional geographic limitations that have always been an impediment.

Companies are eager to feed that consumer hunger for these Internet related goodies. Many are racing to develop the content, applications, and devices they hope will entice more and more consumers to abandon dial-up and slower broadband access in favor of faster broadband. But first, these companies have to be able to reach the broadband consumer.

Thus, usage and deployment of high-speed Internet depends on access to content. Giving broadband consumers the access they want is not a matter of charity; it is a matter of simple good business. Network owners, ISPs, equipment makers, and content and application developers *all* benefit when consumers are empowered to get and do what they wish.

III. MAINTAINING OPENNESS: EMPOWERING CONSUMERS WITHOUT REGULATING THE INTERNET

This is why ensuring that consumers can obtain and use the content, applications, and devices they choose is so critical to unlocking the vast potential of the Internet. Today, broadband consumers generally enjoy such freedom. They can access and use the content of their choice. This easy access includes some of the most promising new uses of broadband. For example, recently the head of the National Cable and Telecommunications Association indicated that cable modem providers would not block traffic from competing Internet voice providers, such as Vonage.[8] Such commitments are good business, but also essential to nurturing competitive innovation.

These general conditions suggest that many, if not most, in the industry recognize that providing access and information is in their own self-interest, particularly as infrastructure providers and developers struggle to discover valuable uses that will enable them to recoup their substantial investments in high-speed architecture. Nevertheless, we must keep a sharp eye on market practices as they continue to evolve and evolve rapidly. And we must do so while safeguarding Congress' intent that the Internet remain free of unnecessary regulation that might distort or slow its growth.[9]

IV. STEERING CLEAR OF POTENTIAL OBSTACLES ON THE HORIZON

Despite the wide-open seas broadband consumers currently enjoy, we must steer clear of obstacles that could appear on the horizon. The high-speed Internet continues to evolve rapidly, and even somewhat unpredictably. Some argue that new threats could undermine consumers' easy use of content, applications, and devices.

Professors Phil Weiser and Joe Farrell make this point in their 2003 paper published with the *Harvard Journal on Law and Technology*.[10] The two professors acknowledge the strong incentives that network owners have to ensure that broadband platforms remain open.[11] Such openness encourages competition among Internet applications and services, which in turn makes platforms more valuable to both consumers and owners.[12] The two note, however, that there may be exceptions to this general rule.[13] They suggest a network owner might face incentives

[8] *See* Donny Jackson, *NTCA: Cable Won't Get In Vonage's Way*, TELEPHONYONLINE.COM (Dec. 19, 2003), *at* http://telephonyonline.com/ar/telecom_ncta_cable_wont.

[9] Telecommunications Act of 1996 § 706(a) (codified at 47 U.S.C. § 157(a)).

[10] Joseph Farrell & Philip J. Weiser, *Modularity, Vertical Integration, and Open Access Policies: Towards a Convergence of Antitrust and Regulation in the Internet Age*, 17 HARV. J.L. & TECH. 85 (2003).

[11] *Id.* at 100-05. A monopolist broadband provider has strong economic incentives to internalize complementary externalities (ICE) by providing access "to its platform when it is efficient to do so, and to deny such access only when access is insufficient." *Id.* at 89.

[12] *Id.* at 103.

[13] *Id.* at 105-19 (The authors outline eight exceptions to the ICE theory which are: (1) Baxter's Law; (2) price discrimination; (3) potential competition; (4) bargaining problems; (5) incompetent incumbents; (6) option value; (7) regulatory strategy; and (8) incomplete complementary.).

to begin restricting uses of their platforms in certain cases: if regulators set prices for using the platform too low,[14] if bargaining among network owners and other companies breaks down,[15] or if companies are just unable or unwilling to recognize their own self-interest in maintaining the freedom broadband consumers want and expect.[16]

This may not be mere academic speculation. There are some troubling restrictions that have appeared in broadband service plan agreements. Professor Tim Wu of the University of Virginia, catalogued some of those restrictions for a symposium here last year.[17] According to his research, these restrictions have included things such as cable companies' early efforts to impose restrictions on virtual private networks, WiFi, and home networking equipment and operation of servers in the home.[18] Although, to the cable companies' credit, many of these concerns have been redressed, press reports continue to plague us alleging that at least some companies have not provided enough guidance to intensive broadband users regarding the limits of their service plans.[19]

The evidence is unclear, however, as to whether and to what degree these restrictions have been aggressively enforced against consumers. Nor is there much evidence that consumers have been denied what they want, even if they are willing to change service plans. Some providers counter that any service plan restrictions have been reasonable attempts to manage their networks to prevent service disruption to customers.[20] Some of the restrictions that have popped up have been removed when it became clear they were not necessary to ensure service quality.

Based on what we currently know, the case for government imposed regulations regarding the use and provision of broadband content, applications, and devices is unconvincing and somewhat speculative. Government regulation of the terms and conditions of private contracts is probably the most fundamental intrusion on the free market. This intrusion is particularly destructive where innovation and experimentation are hallmarks of an emerging service. Such interference should be undertaken only where there is weighty and extensive evidence of abuse.

[14] *Id.* at 105-07.

[15] *Id.* at 112-14.

[16] Farrell & Weiser, *supra* note 10, at 114-17.

[17] Tim Wu, *Network Neutrality, Broadband Discrimination*, 2 J. ON TELECOMM. & HIGH TECH. L. 141, 156-65 (2003).

[18] *Id.* at 159-62.

[19] *See, e.g.*, Associated Press, *Comcast Targets Internet 'Abusers,' But Won't Reveal Limits*, (Jan. 29, 2004), *available at* http://www.securityfocus.com/news/7940.

[20] *See* Wu, *supra* note 17, at 153 (citing Justin Pearse, *UK Shrugs Off American Broadband Troubles*, ZDNET NEWS.COM (Mar. 20, 2000), *at* http://news.zdnet.co.uk/story/0,,t269-s2077792,00.html).

Nonetheless, the industry should take heed of how critical unfettered access to the Net has been, and will continue to be, to the success of broadband. Consumers have a high expectation that such access will continue, and the benefits to them and the nation are significant.

Consequently, it is time to give the private sector a clearer roadmap by which it can avoid future regulation on this issue by embracing unparalleled openness and consumer choice.

V. Consumers Are Entitled to "Internet Freedom"

As we continue to promote competition, we must preserve the freedom of use that broadband consumers expect. Thus, I want to issue a challenge to the broadband network industry to preserve the following "Internet Freedoms."

A. Freedom to Access Content

First, I believe consumers should have their choice of legal content. Consumers expect to be able to go where they want on high-speed connections, and those who have migrated from dial-up would presumably object to paying the premium asked for broadband if certain content were restricted. Thus, I challenge all facets of the industry to commit to allowing consumers to reach the content of their choice. I do recognize that operators have legitimate needs to manage their networks and ensure quality experiences, and reasonable limits sometimes must be placed in service contracts. But such restraints should be clearly spelled out and should be as minimal as necessary.

B. Freedom to Use Applications

Second, consumers should be able to run applications of their choice. As with access to content, consumers have come to expect that they can generally run whatever applications they choose or perhaps even develop. Again, these applications are crucial to continuing the Digital Broadband Migration because they can drive the demand that fuels infrastructure and content deployment. Applications developers must remain confident that their products will continue to work without interference from other companies. No one can know for sure what "killer applications" will emerge to drive deployment of next generation technologies. Again, it is important to challenge all facets of the industry to let the market work and allow consumers to run their applications provided they fall within service plans and will not disrupt the network.

C. Freedom to Attach Personal Devices

Third, consumers should be permitted to attach personal devices they choose to the connections that they pay for in their homes. Devices give consumers more choice, value, and personalization with respect to how they use their high-speed connections, and they are critical to the future of broadband. I challenge all facets of the industry to permit consumers to attach those devices they choose to their broadband connection, so long as the devices operate within their plans, and are not designed and used to enable theft of service.

D. Freedom to Obtain Service Plan Information

Finally, and most importantly, consumers must receive clear and meaningful information regarding their service plans and what the limits of those plans are. Simply put, information is absolutely necessary to ensure that the market is working. Consumers need to know whether and how their service plans protect them against spam, spyware, and other potential invasions of privacy. I challenge all facets of the industry to ensure that consumers can easily obtain this information.

VI. KEY BENEFITS OF PRESERVING "INTERNET FREEDOM"

Numerous benefits follow if industry continues to preserve "Internet Freedom." Internet Freedom will preserve consumers' freedom to access and use content, applications, and devices they choose based on the service plan they choose. It will promote comparison shopping among the growing number of providers by making it easier for consumers to obtain access to meaningful information about the services and technical capabilities they rely on to access and use the Internet.

Internet Freedom promotes innovation by giving developers and service providers confidence to develop applications that will reach consumers and run as designed. Internet voice applications – a notable example that has grabbed the headlines – are all the rage. Internet Freedom ensures that consumers will continue to be able to choose whatever Internet voice service that will function over their high-speed connections.

Preserving Internet Freedom also serves as an insurance policy against the potential rise of abusive market power by vertically integrated providers. If we secure a reasonable balance between the needs of network providers and Internet Freedom, consumers will reap the benefits of broadband *without intrusive regulation*, while preserving industry's incentives to deploy more high-speed platforms.

In closing, I would emphasize that consumers also have a role in this challenge to preserve Internet Freedom. I encourage consumers to challenge broadband providers to live up to these standards and to let the Commission know how the industry is doing. Internet Freedom is intended to give broadband consumers the choices, value, and personalization they are coming to expect and demand. Thus, consumers are the ultimate judges of whether the industry is successful.

I look forward to working with consumers, the industry, and all of you in taking this important step forward in the Digital Broadband Migration. Our voyage continues, but we have begun to see the signs of land. Continuing to keep a sharp watch for dangerous shoals will ensure that someday soon we will dock safely on the shore and begin the bright new day in communications we all hope and dream for.

VII. Question and Answer Session

Following Chairman Michael K. Powell's "Preserving Internet Freedom" speech at the Silicon Flatirons Conference, the Chairman answered questions from students, professors, industry leaders, and journalists. The following is an abridged excerpt of the Chairman's thoughts on recent developments in the telecommunications industry.

A. Was It the Chairman's Intention in Promulgating the Four "Internet Freedoms" to Encourage Service Providers to Post Their Policies and Possibly Make Them Legally Enforceable?

In the last ten years, we have seen the Internet – a phenomenal invention – grow at an unprecedented historical rate which makes it hard to measure its value to consumers and citizens. We bumped into problems along the way. But they are the same kinds of traditional problems that you would expect any new innovation to bump into, whether it be a railroad or a car production line. Either because the government has not really understood it, or, by visionary forethought, the Internet has been left virtually untouched and now thrives. Absent regulation, we have seen some really creative experimentation in various types of self-governance models. Some have failed, but others have succeeded.

As a regulator, it is important to be humble and open minded enough not to assume that just because we have the pen and the authority, we can script out with precision the right way to do this. This is particularly true when there is an information deficit. The Commission does not entirely understand the technical aspects of the services, or how fast they are transforming. The Commission has, however, started to experiment much more with being a catalyst for voluntary initiatives that can avoid regulation, which should be attractive to the industry.

With these four freedoms, providers should become competitive. If I were running a cable company right now, I would love to stand up and say, "Here is what you can expect from Mike's Cable Company, and ask my competitors if they will do the same, and if they do not, then come see Mike's Cable Company." This strategy has worked for wireless local number portability, and it has worked for the do-not-call registry.

There would have to be a demonstrable record of anti-competitive action resulting in consumer harm before regulators should enter into the sacrosanct, private, contractual world between consumer and producer. These types of regulations have a way of getting on the books and never leaving again. Look at the administrative state that we built in the 1930's. The Commission was built on the model of the alphabet agencies of the New Deal. A one-year rulemaking is *moving* in regulatory space, and we are *really* grooving if it is six to seven months. Yet, that is dangerously inadequate for some of these issues. Instead, we would hope that the industry can narrow the number of issues that demand a government response, as opposed to dumping the whole banana under a regulatory sign.

B. How Does the Brand X[21] *Case Affect the Digital Broadband Migration, and How Might the Federal Communications Commission's Strategy Change Depending On the Outcome of the Case?*

I think the case is tremendously significant for the development of broadband policy, not because of the particular result you might ultimately prefer, but because the court will have stolen from the Commission the breadth of discretion that I think it needs to figure out rules of the road in a fast-changing, dynamic capability.

The decision rests on a precedent that is now four years old.[22] At the time the decision was originally reached, just what form and way broadband would flourish was far from clear. Thus, taking discretion to adapt away from the technical expert is dangerous. The substance of the decision would almost say that the Internet has to be a big, fat telephone. In numerous panels and conferences you may debate the minutia of which rules are the right rules and which things are the right things. Most people, however, are beginning to recognize that the Internet and

[21] Brand X Internet Servs. v. FCC, 345 F.3d 1120 (9th Cir. 2003), *en banc rehearing denied*, 2004 U.S. App. LEXIS 8023 (9th Cir. March 31, 2004). The Commission along with the Solictor General are appealing the 9th Circuit's decision to the Supreme Court. *See* Press Release, Federal Communications Commission, Statement of the FCC Chairman on the Government's Appeal of the 9th Circuit's Cable Modem Ruling (Aug. 30, 2004), *available at* http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-251527A1.doc.

[22] *See* AT&T Corp. v. City of Portland, 216 F.3d 871 (9th Cir. 2000).

its potential is something farther reaching and greater than the facile mind that wants to embrace it, call it a big, fat telephone, and then pile on 100 years of telecommunication regulation. Much of that regulation, by the way, has never been thought through, or filtered for its relevance or applicability to a new, emerging service.

Alternatively, if the federal government or state governments want to plow through the technical realities of the Internet, its potential, and its benefit to consumers, and slowly come to the same conclusion over the next hundred years that it ought to be regulated that way, then fine. But that is not what you would be doing here. You would, in almost a lazy move, be extrapolating rules that have built up around an entirely different network, an entirely different economic model, and an entirely different role for regulation that would not reflect any of the more enlightened and far-reaching thinking. Shame on us if we do. We will be wondering why we are 30th in world in broadband deployment, leading to more outsourcing of jobs outside the country, and more productivity losses in the United States.

C. How Does Regulating Broadcasters' Content Competitively Impact the Cable Companies Vis-à-Vis the Network Companies?

There is no area of passionate public discourse understood less than this one. The indecency statutes that are on the books have only been applied to the broadcast, free, over-the-air medium. There are a number of legal, intellectual, and constitutional reasons why that is the case. First, there is an assumption that broadcasting uses the public's property for free. In exchange, the broadcasters have a higher public trust obligation. That has been the government's broadcasting model for seventy years. That rationale is becoming more tenuous because 88% of Americans subscribe to cable or DBS,[23] and increasingly the Internet, and increasingly Xbox, and increasingly Blockbuster Video, and increasingly XM satellite radio. Our society is being bombarded from multiple avenues with media, information, and entertainment. I think you start to have wobbliness in the outlook of the government if it is always myopically focused on one segment – here by the way, the more declining media sphere – and that is the way the statutes are currently applied.

The Supreme Court has said that with respect to free, over-the-air broadcasting, the government can go further than it normally can regulating other media outlets. It has a lesser First Amendment standard

[23] Annual Assessment of the Status of Competition in the Market for the Delivery of Video Programming, *Tenth Annual Report*, 19 FCC Rcd. 1606, 1609-10, at ¶ 7 (2004).

than newspapers, for example. The Court has also said that cable has a First Amendment interest somewhat akin to that of newspapers, as opposed to broadcasting.[24] One reason is that there is a subscription relationship. You can have it or not, and the consumer has expressed a voluntary accession to the medium. Secondly, much of the programming that you talk about has another level of filter in the sense that you have to subscribe to HBO.

These are not my parameters, but these are the ones that the courts have employed that limit the restrictions. The bottom line is the same. At some point, if the country is serious about wanting to debate what the public interest is in the media, then it is going to have to broaden its mind and its perspective enormously. I am going to use my children as an example: ask them if they know what a broadcast channel is. They do not. They have a clicker in their hand and it goes 7, 9, 10, 12, 159, 222, and they do not know the difference between 214 and 7. I find it phenomenal that the First Amendment changes channels too.

The Commission is aggressively enforcing the law as written. I think that if you want to talk about the effect of these mediums in our society, you are really kidding yourself if you think you can wall off one small part so your children never hear the "F-word" again through other mediums. Regulating what our children watch is an important issue, as is how to balance the role of parental control versus government control. That is always a healthy thing to talk about, but if you notice, boys do not watch TV. The recent Neilsen studies are shocking in displaying the degree to which little boys have left the television space in big numbers.[25] Why? They live on their video machines. Have you seen Grand Theft Auto?[26]

D. Would It Be an Abuse of Market Power for Cable Modem Providers to Offer Limited Internet Access for a Lower Price to Individuals Not Desiring Complete Access?

No one is talking about not allowing network providers to enter the applications development business. We are not asking network providers to be dumb wholesalers with no other ability to provide access to high value services. I think that we would kill them if we did that. The economics would make it very difficult to do that unless we started doing what we did in 1913, and start re-embracing monopolies in order to

[24] United States v. Playboy Entm't Group, Inc., 529 U.S. 803, 815 (2000).

[25] *See, e.g.,* Joe Mandese, *Video Games Emerge As "No. 4" Medium, Displace Print Among Young Guys*, MEDIA DAILY NEWS (Apr. 5, 2004), *available at* http://www.mediapost.com/dtls_dsp_news.cfm?newsID=245176.

[26] Grand Theft Auto is a product of Rockstar Games. ROCKSTAR GAMES, INC. *at* http://www.rockstargames.com (last visited Aug. 14, 2004).

guarantee their rates of return for the developments that we want. I think that we do want them to go in that space, and, by the way, some of them are good at it. I have this running fight with my mother-in-law, and I will not disparage any product, but she pays a lot for a particular e-mail provider, and I cannot get her to switch to save my life. Given the way she uses it, she would actually be better off somewhere else. But she is very comforted by it, and she is willing to pay for it. To deny her that is also to deny her choice.

There is a problem that we have that I call techno-ecstasy. We think that because we can do cool things – everyone should, but not everybody wants to do that. There is a reason we have editors, aggregators, and simplifiers. A lot of people want someone to bring order to a chaotic world, and I think there is going to be room for that. I think that everybody that wants to offer that should be able to do so. You just have to be careful that in your zest to offer it, you are also willing to knock the gates over for everyone else.

E. Have New Technologies Such as Voice Over IP, and Wireless Broadband Made Universal Access a Thing of the Past?

I have said before that Universal Service is an objective, and the objective is ubiquity and affordability. It is our commitment to give every American access to tomorrow's technology at affordable rates. There is nothing about that noble goal that is any less compelling today with advanced technologies, than it was yesterday. My only argument is that you ought to be very creative and thoughtful about how you use different technologies to solve the problems, rather than just lightly assuming that you have to approach the problems the exact same way you approached them for the last one hundred years.

In 1913 MaBell sold us this bunch of goods – let us be a monopoly and we will do it for you. It was not a game – it worked – and that is because every single hamlet and town and mountaintop was going to be reached the same way, and that way was prohibitively expensive in large parts of the country. I suggest that we start to get technologically savvy about solving rural universal service.

Universal service is so hard in rural America because it is hard to string a twisted copper cable 600 miles up the side of a mountain. Instead, contrary to many of the pundits, what we see is an explosion of wireless innovation in rural America. This year I have been to several very rural areas where they are employing wireless Internet services, usually by guys who buy the equipment at Circuit City and put it in the barn, put an antenna on, and come up with a community solution for subscription. All of a sudden, they have better broadband than we do in Washington. It is amazing what a little room will do for an innovator.

The economics are fundamentally different, and that gives me a lot of hope and optimism that we can solve problems in parts of the country using a different approach than what we might have used before. None of this suggests the final outcome, only that I challenge those in Washington, and policymakers among states, to be a little more aggressive in thinking about how to take care of our rural citizens in the future with technology in our approaches, instead of assuming that everything is a really huge pot of money that must be used in exactly the same way, in exactly the same form, when you are in different places.

We should take the lessons learned in the unlicensed spectrum. We need to get it out of the hands of legal thinkers and into those of technical innovators and say here is your driver's license, do not speed, and do not break the law. I do not care whatever else you do, I do not care what color car it is, what size it is, or what shape it is. Just follow certain rules that prevent interference meltdown, and feel free to figure out what to do with the spectrum for your community, rather than us at our command and control computers deciding what will be used for what purpose on this hour of this day. I think the challenge is to make better use of the spectrum that we have currently licensed, allocate more unlicensed spectrum, and allow greater flexibility so that rural constituents can take advantage of opportunities in their area without having to deal with a heavy regulatory council constantly arguing about the right way to do it.

F. Would It Be Necessary to Regulate to Prevent Americans From Getting Illegal Content?

You should not regulate, you should *prosecute*. As deregulators, we have not gotten so absurd to think that free markets should allow murder. There is always an important distinction between permitting legal conduct, and facilitating illegal content. The kinds of free market values that are important are ones that are faithful to a rule of law. It is not about doing whatever you want. The market is fundamentally a dialogue between the producer and the consumer. They have a dialogue about what they want, at what price, and about what they will be willing to accept and what they will not. They did not invite the regulators to the dinner table, and regulators should not accept an invitation to the dinner table unless there are clear, demonstrable reasons to be there. Regulators cannot interfere with a relationship that produces ultimate welfare, at least as shown by the history of economics as we have come to know it. Many systems in the history of the world have attempted to do it better, but I have not seen the one that actually does it. It is shocking to me that every other decade you have to re-win the argument that five smart people sitting in Washington cannot micro-plan the economy.

There are other kinds of law. There is a big difference between regulation, property rights, and contractual protections. There is no such thing as a free market without a rule of law that is fully enforced, and carefully protected. To return to the question, that is why copyright law is so important. The current debate is about what will be the legitimate property interests of holders in exchange for allowing your viewing or listening. That is a fundamental property right notion that the government sanctions, but it is about private property, not public confiscation. I love the music stuff, legally.

G. What Happens When the Legal Regimes Fail? We Have Laws That Prevent Downloading Free Music, But That Does Not Seem to Have Stopped It?

First of all here is a warning to all who will be producing products in the Internet space: you better watch these kids, because you are beginning to see in their hands the tools to solve problems if you will not solve them first. I watch my children think differently than I would have thought to do in the same situation. They look for something first for only a little while. If it is not there, they start figuring out how to make it.

Sean Fanning wanted something the music industry was unwilling to give him, so he made it. There is going to be a premium on those who wish to take your money to be faster and more responsive to the digital generation's needs. They do not have much patience before they start solving problems themselves. Producers can provide compelling value as well as legal deterrents, so that most people, the ones that matter, conduct themselves legally. When Apple iTunes figured out how to do what it did, and it is up to nearly 100 million downloads,[27] there was a noticeable decrease in illegal downloading. While it has not been extinguished, it is also still in the very early innings of trying.

You have to worry about losing a generation that has become very acclimated to something. I am puzzled by the behaviors of my fourteen year old's generation. They will not hesitate to spend a fortune on a video game, which as constructed is not that much more expensive to make than the music they seek to download for free. And the same kids that are downloading music for free are paying a $1.59 to listen to crappier music on their phones. Ringtown downloading was a $3.5 billion industry.[28] This is all about winning the hearts and minds, and

[27] Laurie J. Flynn, *IPod Demand Leads Big Increase in Earnings for Apple*, NY TIMES, July 15, 2004, at 4.

[28] Reuters, *Ring Tones Bringing in Big Bucks*, WIRED NEWS (Jan. 13, 2004), *available at* http://www.wired.com/news/business/0,1367,61903,00.html.

acclimating people to a direction. Producers will have to be swifter and more in tune at an earlier stage with consumers.

H. Do You Think it's Time for a New Telecom Act and If So, What Should Be In It?

What I will give you is food for thought: the world of communications is very different now, after the 1996 Act, than it was before. The administrative philosophy behind the 1934 Act is one of enormous delegation of authority with very flexible standards, and little determinacy. The flexibility in the Act was based on trust in the agency as being an enlightened group of individuals. There is this great book by professor James Landis on the theory of administrative law,[29] which states that we will staff these commissions with wise people of special intellect. The notion was that we would invest in these special people, these obligations, and then we would give great deference to them and that is the way it would work.

The 1996 Act is very different. It attempts to be a comprehensive blueprint about every intricate question and it tries to put very serious restraints on the Commission. It is 750,000 words long; every Senator I have met swears they personally wrote it. It is fraught with ambiguity and inconsistency. It is a very, very different kind of model, and the interesting question is which – '34 Act, or '96 Act – would be better for the Internet world?

The difference between the two is similar to the question of whether you believe the Constitution is about original intent, or if you believe it has to be a living constitution. I do not know what I think about constitutions, but in this space it has to be living regulatory policy. We have to go back to thinking about what the shear structural and intellectual limitations of regulatory authority are, as well as the shear immensity of what is unknown as opposed to what is known, and finally, build an institution and a law that is capable of being dynamic and adaptive, as opposed to oppressive and proscriptive. You have to go back to principles and standards – a really smart institution that has technological expertise to learn, change, adapt, and not be so parochial. When an agency becomes proscriptive, it invites an enormous amount of litigation.

I am convinced that there is a formula such that for every one word of law, you have multiplied by ten the number of lawsuits possible. The Telecom Act of '96 will never settle – every company can find the words they are looking for to fight decisions. There is also a lot of political mischief within the Commission. Somebody can come in and say: "I've

[29] JAMES M. LANDIS, THE ADMINISTRATIVE PROCESS (1938).

got a way that you can do this – there's this one word in footnote nine that I've stuck in there."

I tried to figure out how I should manage the Commission knowing that we do not know all the answers. I remembered watching a basketball game with Duke having just successfully beaten another team. I was shocked by the degree to which each player, in an unbelievably disciplined way, could mimic the basic philosophy of the coach. When they step on the court, they do not know what is going to happen. The players are an organism and they are built to read, adjust, and adapt. The discipline of their organization and the proficiency in skills allows them to do that.

The same is true for soldiers. I was a soldier, and we cannot train soldiers. We cannot tell them what will happen on a battlefield. You condition them to be skillful, disciplined, creative, innovative, and adaptive and you trust them. You trust them to make adjustments and adapt and persevere through change because no two battles are the same.

The Commission and developing policy must be thought of in the same way. It has to be a dynamic living organism. And that is why I am a fervent advocate of free markets, not only because I think they are superior but because it is the model that says restrain yourself, be humble about what you do not know.

WOULD MANDATING BROADBAND NETWORK NEUTRALITY HELP OR HURT COMPETITION?

A COMMENT ON THE END-TO-END DEBATE

CHRISTOPHER S. YOO[*]

ABSTRACT

A chorus of commentators has drawn inspiration from the "end-to-end" argument first advanced by Saltzer, Reed, and Clark and called upon policy makers to mandate that last-mile broadband providers adhere to certain principles of network neutrality. In his contribution to this symposium, Professor Christopher Yoo offers an economic critique of these proposals. He first concludes that they are based on a misreading of Saltzer, Reed, and Clark, who implicitly reject turning the end-to-end argument into a categorical mandate. In addition, prohibiting the use of proprietary protocols can harm consumers by skewing the Internet towards certain types of applications. Finally, network neutrality raises the even more significant danger of forestalling the emergence of new broadband technologies by reinforcing the existing supply-side and demand-side economies of scale and by stifling incentives to invest in alternative network platforms. Although such considerations would be problematic under any circumstances, they carry particular weight with respect to industries such as broadband, which are undergoing rapid technological change.

[*] Associate Professor of Law, Vanderbilt University. Thanks to the participants on the panel on Broadband Policy at the Conference on "The Digital Broadband Migration: Toward a Regulatory Regime for the Internet Age" sponsored by the Silicon Flatirons Telecommunications Program at the University of Colorado and to Douglas Galbi, Richard Nagareda, Bob Rasmussen, Doug Sicker, Jim Speta, Phil Weiser, and Tim Wu for their input on earlier drafts of this paper.

INTRODUCTION

The broadband industry has reached a crossroads. After avoiding the issue for years,[1] the Federal Communications Commission (FCC) has opened two comprehensive proceedings designed to resolve how the major broadband technologies should be regulated.[2] Congressional committees have also conducted hearings exploring many of the same issues.[3] At the same time, a chorus of commentators, led by Stanford law professor and Internet guru Lawrence Lessig, has invoked the "end-to-end argument" first advanced by Jerome Saltzer, David Reed, and David Clark in 1981[4] and has called upon the FCC to require that all broadband network owners adhere to certain principles of open access and network neutrality.[5] At their core, network neutrality proposals stem

[1] The FCC's reluctance to address these issues may end up limiting its latitude in determining how broadband should be regulated. When the Ninth Circuit first confronted the proper regulatory classification of cable modem services in *AT&T Corp. v. City of Portland*, 216 F.3d 871 (9th Cir. 2000), the FCC had not yet addressed the issue, *see id.* at 876, which forced the court to resolve the issue for itself by concluding that cable modem service is a "telecommunications service." Even though the FCC has since concluded that cable modem service is more properly regarded as an "information service," the Ninth Circuit has declined to accord *Chevron* deference to the FCC's rulings on the grounds that it is bound by stare decisis to adhere to its initial determination. *See* Brand X Internet Servs. v. FCC, 345 F.3d 1120 (9th Cir. 2003). This appears inconsistent with *Chevron*'s recognition that agency interpretations of statutes should be permitted to change over time. *See* Chevron USA Inc. v. Natural Res. Def. Council, 467 U.S. 837, 863-64 (1986). For an interesting discussion of the relationship between *Chevron* and stare decisis, see Richard L. Pierce, Jr., *Reconciling* Chevron *and Stare Decisis*, 85 GEO. L.J. 2225 (1997).

[2] One docket addresses the regulatory regime to be applied to digital subscriber line (DSL) service. *See* Appropriate Framework for Broadband Access to Internet over Wireline Facilities, *Notice of Proposed Rulemaking*, 17 FCC Rcd. 3019 (2002) [hereinafter *Wireline Modem NPRM*]. The other docket focuses on cable modem services. *See* Inquiry Concerning High-Speed Access to the Internet Over Cable and Other Facilities, *Declaratory Ruling and Notice of Proposed Rulemaking*, 17 FCC Rcd. 4798 (2002) [hereinafter *Cable Modem Declaratory Ruling and NPRM*]; Inquiry Concerning High-Speed Access to the Internet Over Cable and Other Facilities, *Notice of Inquiry*, 15 FCC Rcd. 19,287 (2000) [hereinafter *Cable Modem NOI*].

[3] *See The Government's Role in Promoting the Future of the Telecommunications Industry and Broadband Deployment: Hearings Before the Sen. Comm. on Commerce, Science and Transportation*, 107th Cong., 2d Sess. (2002).

[4] *See* J. H. Saltzer et al., *End-to-End Arguments in System Design*, 2 ACM TRANSACTIONS ON COMPUTER SYS. 277 (1984) (revised version of paper first presented in 1981).

[5] Lessig offered his most complete statements of this position in LAWRENCE LESSIG, THE FUTURE OF IDEAS 34-48, 147-75 (2001); and Mark A. Lemley & Lawrence Lessig, *The End of End-to-End: Preserving the Architecture of the Internet in the Broadband Era*, 48 UCLA L. REV. 925 (2001). For other leading commentaries offering related proposals, see Jim Chen, *The Authority to Regulate Broadband Internet Access over Cable*, 16 BERKELEY TECH. L.J. 677 (2001); Mark Cooper, *Open Access to the Broadband Internet: Technical and Economic Discrimination in Closed, Proprietary Networks*, 71 U. COLO. L. REV. 1011 (2000); William P. Rogerson, *The Regulation of Broadband Telecommunications, the Principle of Regulating Narrowly Defined Input Bottlenecks, and Incentives for Investment*

from the concern that network owners will use their control over last-mile broadband technologies to discriminate against nonproprietary Internet service providers (ISPs) and unaffiliated content and applications. According to these advocates, mandating open access interoperability is essential if the environment for competition and innovation on the Internet is to be preserved.

There can be no question that the widespread acceptance of the end-to-end argument has played a key role in fostering the Internet's meteoric success and remains a central tenet guiding decisions with respect to network design. That said, the academic debates and the arguments currently being advanced before the FCC have largely overlooked the fact that there is a crucial difference between embracing the end-to-end argument as a *design principle* and elevating it into a *regulatory mandate*. While adherence to the end-to-end argument may make sense in most cases, circumstances do exist in which mandating network neutrality would actually harm competition.

In this article, I develop three fundamental propositions that shed new light on the end-to-end debate. The first is that the leading network neutrality proposals are actually inconsistent with the end-to-end argument advanced by Saltzer, Reed, and Clark. A close reading of their seminal works supports applying the end-to-end argument on a case-by-case basis rather than in the categorical manner envisioned by the current proposals pending before the FCC, a conclusion confirmed by subsequent technologists.

Second, I show how network neutrality proposals in essence are rooted in concerns about vertical integration. Application of the conventional economic wisdom about vertical integration reveals that the dangers envisioned by network neutrality advocates are likely to be more imaginary than real. Although considerable disagreement exists over many aspects of vertical integration theory, there is widespread agreement that certain structural preconditions must be satisfied before vertical integration can plausibly threaten competition. An empirical analysis reveals that these preconditions are not met with respect to the broadband industry.

Third, I would like to outline a new economic approach that offers a radically different approach to promoting competition in the physical layer. One of the core insights of vertical integration theory is that any chain of production can maximize economic welfare only if every level of production is competitive. In other words, any chain of production is

and Innovation, 2000 U. CHI. LEGAL F. 119; Daniel L. Rubinfeld & Hal J. Singer, *Open Access to Broadband Networks: A Case Study of the AOL/Time Warner Merger*, 16 BERKELEY TECH. L.J. 631 (2001); Timothy Wu, *Network Neutrality, Broadband Discrimination*, 2 J. ON TELECOMM. & HIGH TECH. L. 141 (2003).

only as efficient as its least competitive link, which in the case of the Internet is undoubtedly the last mile. In attempting to preserve and encourage competition and innovation in applications, content, and ISP services, these proposals are directed towards increasing competition in those segments of the broadband industry that are already the most competitive. Instead, basic economic principles suggest that the better course would be to eschew attempting to foster competition in ISP services, content, and applications and instead to pursue regulatory options that would promote competition in the segment that is most concentrated: last-mile technologies.

Restated in terms of the existing models of "layered competition,"[6] the major network neutrality proposals advocate regulating the logical layer in a way that promotes competition in the application and content layers. In the process, they direct their efforts towards the wrong policy problem. Instead, the focus of public policy should be to promote competition in the physical layer, which remains the level of production that is currently the most concentrated, the least competitive, and best protected by barriers to entry.

Finally and perhaps most importantly, the standardization implicit in compelled interoperability tends to reinforce and entrench the sources of market failure that have historically limited the level of competition among last-mile technologies. The traditional justification for regulating wireline communications networks is that the presence of large, up-front sunk costs creates large supply-side economies of scale that cause markets for telecommunications services to collapse into natural monopolies. Interestingly, allowing networks to differentiate the services they offer can mitigate the tendency towards natural monopoly by allowing multiple last-mile technologies to coexist notwithstanding the presence of unexhausted returns to scale. Providers confronting cost disadvantages inherent in the smaller scale of their operations can survive by tailoring their networks to the needs of subgroups who value a particular type of network services particularly highly in much the same manner that specialty stores survive in a world dominated by one-stop shopping. Permitting variations in the protocols and network infrastructure employed in broadband networks thus might enable competition to exist notwithstanding the presence of unexhausted returns to scale.

For example, it is conceivable that allowing networks to differentiate themselves might make it possible for multiple last-mile networks to coexist by serving the needs of a different subgroup: one optimizing its network for conventional Internet applications such as e-mail and website access, another incorporating security features to facilitate e-

[6] *See infra* Section III.A.2.

commerce, a third employing routers that prioritize packets in the manner needed to facilitate time-sensitive applications such as Internet telephony, generally known as "voice over Internet protocol" (VoIP), with others targeting other needs. Conversely, mandating interoperability commodifies bandwidth in ways that sharply limit opportunities to compete on dimensions other than price, which reinforces the advantages enjoyed by the largest and most established players. Moreover, by favoring innovation at the network's edge to the exclusion of innovation in the network's core, this approach risks introducing a regulation-induced bias in favor of certain types of applications and against others.

Other commentators have invoked the burgeoning literature on network economic effects as an alternative justification for regulatory intervention.[7] Network economic effects exist when the value of network access depends on the number of other users connected to the network, rather than the network's technological characteristics or price. The more people that are part of the network, the more valuable the network becomes. As a result, a user's decision to join a network increases the value of the network for others. The fact that the new user cannot capture all of the benefits generated by their adoption decision has led many theorists to regard network economic effects as a kind of externality that causes overall network utilization to drop below efficient levels. Some commentators also argue that network externalities can turn network access into a competitive weapon. By refusing to interconnect with other networks, network owners can force users to choose one network to the exclusion of others. Forcing users to commit to one network naturally leads users to flock to the largest network. In short, network economic effects can create demand-side economies of scale analogous to the supply-side economies of scale caused by the presence of sunk costs.

The current debate has overlooked a number of critical considerations that make it implausible that network economic effects are likely to harm competition.[8] Even more importantly for the debates

[7] *See, e.g.*, Jerry A. Hausman et al., *Residential Demand for Broadband Telecommunications and Consumer Access to Unaffiliated Internet Content Providers*, 18 YALE J. ON REG. 129 (2001). For the seminal works in the theory of network economic effects, see Joseph Farrell & Garth Saloner, *Standardization, Compatibility, and Innovation*, 16 RAND J. ECON. 70, 70 (1985); Michael L. Katz & Carl Shapiro, *Network Externalities, Competition, and Compatibility*, 75 AM. ECON. REV. 424 (1985).

[8] As I will subsequently discuss in greater detail, the theory of network externalities are largely inapplicable to physical networks such as telecommunications networks, since the network owner is in a position to internalize whatever externalities that may exist. *See infra* notes 115-117 and accompanying text. Furthermore, a network must possess market power before network economic effects can even plausibly harm competition. *See infra* notes 113-114 and accompanying text. As I discuss in *infra* Section III.B.1, this precondition is not

surrounding network neutrality, compelled standardization runs the risk of reinforcing the tendencies towards concentration already extant in the broadband industry. The economic literature recognizes that network diversity can ameliorate the anticompetitive effects of the demand-side economies of scale associated with network economic effects in much the same manner as it can mitigate the problems caused by supply-side economies of scale. Imposing network neutrality would prevent such competition from emerging and would instead force networks to compete solely in terms of network size and price, considerations that tend to favor the largest players. As a result, imposing network neutrality as a regulatory matter can have the perverse effect of entrenching the oligopoly of last-mile providers that represents the central policy problem facing the broadband industry. In other words, mandating network neutrality raises the real danger that regulation would become the source of, rather than the solution to, market failure. Such considerations are particularly problematic when the industry is undergoing dynamic technological change, as is the case in broadband.

Emphasizing the potential harms associated with compelling network neutrality as a regulatory matter is not inconsistent with recognizing the value of adhering to standardization as a default principle. Interoperability and the end-to-end argument clearly offer benefits to both providers and consumers, and network designers should hesitate before deviating from those central precepts. Indeed, I would expect that most industry participants would voluntarily design their technologies to be fully interoperable and compatible in the vast majority of cases even in the absence of regulation. The question posed by the debate over network neutrality is not whether consumers benefit from standardization; they clearly do. To the extent that is true, there is no need to mandate network neutrality, since the benefits to consumers from standardization should be reflected in market outcomes. The real issue posed by the network neutrality debate is whether regulators should step in and impose standardization in those situations where the market exhibits a preference for differentiation. The fact that the structure of the broadband industry makes it unlikely that any network owner will be able to use nonstandardization to harm competition indicates that such intervention is unwarranted. In addition, by preventing last-mile providers from tailoring their networks to pursue alternative strategies, barring network diversity threatens to make matters worse.

The balance of this article is organized as follows. Section I describes the Internet's basic structure and lays out the issues surrounding the network neutrality debate. Section II evaluates the end-to-end

satisfied with respect to the broadband industry.

argument, concluding that it does not support the imposition of network neutrality as a regulatory mandate. Section III demonstrates the close relationship between network neutrality and the economics of vertical integration. It also examines the structure of the broadband industry, concluding that the structural preconditions needed for vertical integration to pose a threat to competition are not satisfied. Section IV analyzes the potential benefits of allowing last-mile providers to deviate from complete interoperability. Allowing last-mile providers to use vertical integration to differentiate their networks would allow the realization of certain efficiencies and would permit them to offer a broader range of services better attuned to consumers' preferences. Even more importantly, requiring all broadband networks to use nonproprietary protocols can actually reduce competition by reinforcing the economies of scale already enjoyed by large telecommunications providers. Section V discusses the proper role of regulation, concluding that regulatory authorities will be more effective at promoting entry by new network platforms than they would be in ascertaining whether a particular exclusivity arrangement would promote or hinder competition. Indeed, one of the benefits of pursing the strategy of promoting entry is that it has embedded within it a built-in exit strategy. Once a sufficient number of broadband network platforms exist, regulatory intervention will no longer be necessary.

I. FRAMING THE NETWORK NEUTRALITY DEBATE

Understanding the debates about broadband regulation requires an appreciation for certain key features of the Internet's underlying structure.[9] In order to facilitate the discussion, Part A offers a simplified description of the basic structure of the original narrowband Internet. Part B identifies the key architectural changes effected by the migration to broadband technologies. Part C considers the impact of shifts in users' relationship with the Internet. Part D examines how these various transformations have shaped the debates about network neutrality that have arisen in the broadband regulatory proceedings.

A. The Architecture of the Narrowband Internet

As has been often noted, the Internet is not a single, monolithic network. Rather it is a network of networks that are interconnected together. When the Internet first became popular, it was fairly easy to divide components of the network into three categories.[10] The core of

[9] Those already familiar with the Internet and the debates about network neutrality may wish to skip directly to Section II.

[10] *See* Inquiry Concerning the Deployment of Advanced Telecommunications

the Internet is provided by backbone providers, such as AT&T, Cable & Wireless, Level 3, MCI WorldCom, and Qwest.[11] Backbones are high-bandwidth, long-haul network providers that carry traffic between a limited number of recognized locations. By 1998, backbones interconnected through eleven public access points.[12] Since that time, major backbone providers have increasingly interconnected directly at private locations.

The final connection is provided by last-mile providers, which carry data traffic from central facilities located in different metropolitan areas to end users. In the narrowband world, last-mile services are almost invariably provided by the local telephone company. Narrowband customers typically connect by using a dial-up modem to place a conventional telephone call routed to another location within the same local calling area. Customers with higher volumes of data traffic employ more sophisticated telephone technologies, such as T-1 or T-3 lines, integrated services digital networks (ISDN), frame relay, or fiber optics.[13]

The gap between the limited geographic points served by backbone providers and the widely dispersed locations of last-mile providers is bridged by a third category of network provider, commonly called ISPs.[14] The best known ISPs include America Online, MSN, Earthlink, Juno, and Netzero. ISPs typically have a higher port density and carry a lower volume of traffic at lower speeds than backbone providers. In addition to carrying traffic between the NAPs and the points of presence

Capability to All Americans in a Reasonable and Timely Fashion, and Possible Steps to Accelerate Such Deployment Pursuant to Section 706 of the Telecommunications Act of 1996, *Second Report*, 15 FCC Rcd. 20,913, 20,922-38 ¶¶ 16-59 (2000) (categorizing Internet network providers into a similar three-part taxonomy).

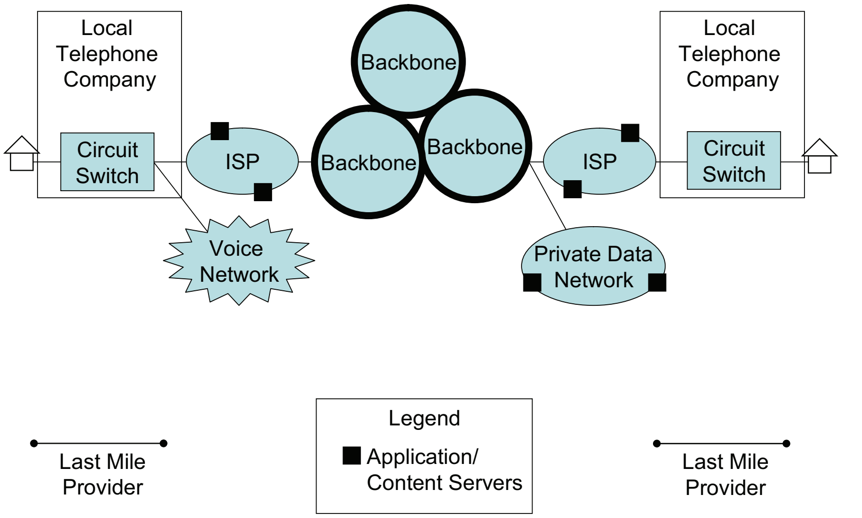
[11] Backbone providers are also called "tier 1 ISPs."

[12] The original backbone supported by the National Science Foundation until 1995 (known as NSFNet) carried traffic between three "network access points" (NAPs) located in San Francisco, Chicago, and New York. The restrictions the NSF placed on commercial uses of the backbone led a group of private companies to create an additional interconnection point known as the "commercial internet exchange" (CIX) located in Santa Clara, California. The federal government also established federal internet exchange (FIX) points in College Park, Maryland, and Mountain View, California. In addition, Metropolitan Fiber Systems, Inc. (now owned by WorldCom) expanded the fiber rings that it established in Chicago, Dallas, Los Angeles, San Jose, and Washington, D.C. into "metropolitan area exchanges" (MAEs) that essentially performed the same functions as NAPs. *See* Jack Rickard, *The Internet-What Is It?*, BOARDWATCH, Winter 1998, *available at* http://www2.cs.uh.edu/~klong/papers/WhatIsTheInternet.pdf; Michael Kende, *The Digital Handshake: Connecting Internet Backbones*, 11 COMMLAW CONSPECTUS 45, 48-50 (2003).

[13] Kende, *supra* note 12 at 46.

[14] National companies who connect local points of presence to NAPs are often called tier 2 ISPs. Regional providers are often called tier 3 ISPs. Note that many providers that I have termed backbone providers refer to themselves as ISPs. For simplicity, I will refer to tier 1 ISPs as "backbone providers" and reserve the term "ISP" for tier 2 and tier 3 providers.

FIGURE 1
BASIC ARCHITECTURE OF NARROWBAND TECHNOLOGIES



within each last-mile provider's service area, ISPs perform a number of other functions, including supplying e-mail servers, hosting end users' webpages, and caching the most popular content locally so that customers can access it more easily. ISPs also often offer portal services and proprietary content, which allow them to add value through their "unique aggregation and presentation of content that allowed for easy consumption by end users."[15]

Once a narrowband ISP receives a call, the ISP demodulates the signal from the dial-up modem and routes the traffic onto its own packet-switched networks. If the packets are addressed to a destination located on the same ISP network (such as an e-mail address associated with a different customer of the same ISP), the ISP conveys them to their destination without involving any other ISPs or backbones. If the packets are addressed to a more distant location, the ISP hands off the packets to a backbone provider, which in turn may hand off the packets to one or more downstream backbone providers. Eventually, one of the backbones hands off the packets to the destination ISP or a private data network, which in turn delivers them to their termination point.

The narrowband network configuration possesses two features that have influenced the debates about network neutrality. First, the last-mile

[15] Rubinfeld & Singer, *supra* note 5, at 634.

provider does not need to maintain any packet-switching capability of its own. Instead, it simply routes calls it receives on an inbound local telephone line through its central office switch to an outbound local telephone line without modifying the traffic in any way. This transparency makes last-mile narrowband connections nondiscriminatory. Because customers can use the local telephone network to call any other customer connected to the network, all a narrowband ISP needs from the last-mile provider is an appropriate number of incoming business lines.

Second, the movement of packets through ISPs and backbone providers is controlled by a family of nonproprietary protocols known as the transmission control protocol/Internet protocol (TCP/IP). For our purposes, the most distinctive feature of TCP/IP is that it routes all packets in a nondiscriminatory (i.e., first come, first served) manner without regard to the packet's content, point of origin, or associated application.

B. Architectural Changes Resulting from the Migration to Broadband

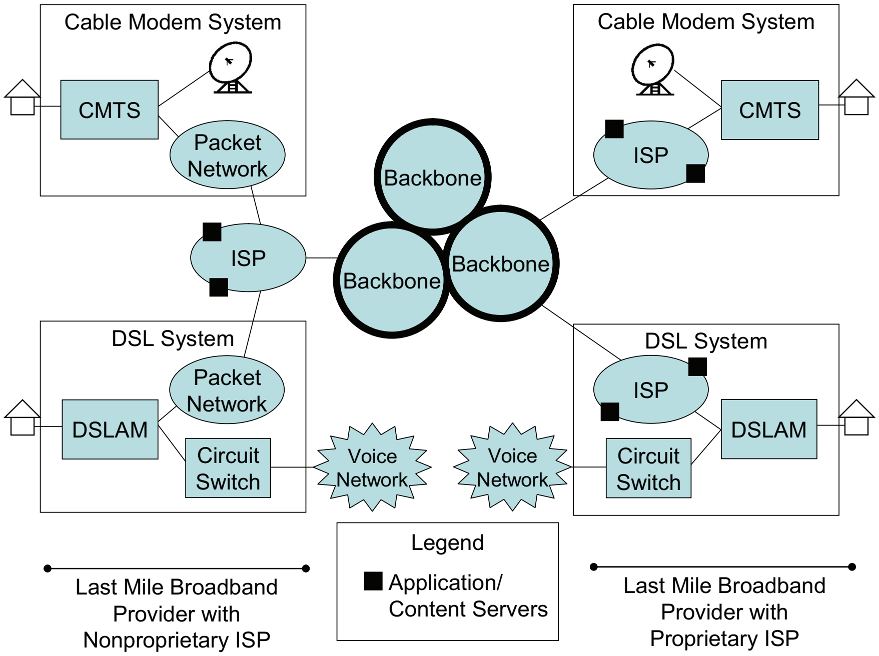
The arrival of broadband technologies has effected some fundamental changes in the Internet's architecture. Many residences and small businesses now have the option of contacting the Internet through cable modem systems maintained by local cable operators, such as Comcast, Time Warner Cable, Cox, and Charter, or through a digital subscriber line (DSL) service offered by local telephone companies, such as Verizon, SBC, Qwest, and BellSouth.

Because both DSL and cable modem providers use the same infrastructure to provide two different types of service (either cable television combined with cable modem service or local telephone service combined with DSL), both types of providers must maintain equipment to segregate the two different communication streams. DSL systems route traffic through devices known as a digital subscriber line access multiplexers (DSLAMs), which separate the voice communications from the data-based communications.[16] Cable operators employ devices known as frequency up-converters and a cable modem termination systems (CMTSs) to divide the video and data streams.[17]

[16] Note that although most DSLAMs are located in the central office switch, some local telephone companies are deploying digital loop carrier (DLC) architectures that allow DSLAMs to be located in remote terminals. Locating DSLAMs closer to end users represents one way to increase the coverage area of DSL service. *See* Daniel F. Spulber & Christopher S. Yoo, *Access to Networks: Economic and Constitutional Connections*, 88 CORNELL L. REV. 885, 1004-05 (2003).

[17] *See id.* at 1014-15.

FIGURE 2
BASIC ARCHITECTURE OF BROADBAND TECHNOLOGIES



Unlike what was the case in the narrowband world, last-mile broadband providers must maintain a packet-switched network in their main facilities to hold and route the stream of data packets after they have been separated from other types of communications. Thus, under a broadband architecture, last-mile providers no longer serve as mere pass-throughs. They must instead necessarily perform the same routing functions previously carried out by ISPs. Indeed, some last-mile broadband providers have negotiated their own interconnection agreements with backbone providers and require all of their customers to use their own proprietary ISP, thereby supplanting the role of independent ISPs altogether. The migration of Internet users from narrowband to broadband technologies has thus had the inevitable effect of reducing the viability of many independent ISPs and encouraging last-mile providers to bundle their offerings with ISP services.

C. Shifts in User Demand

The advent of broadband technologies has also largely coincided with a number of fundamental changes in user demands that are placing increasing pressure on the continued adherence to a uniform, TCP/IP-based architecture. Although the forces are somewhat complex, a few examples illustrate the forces driving this fundamental shift.[18]

1. The Shift from Institutional to Mass-Market Users

The termination of NSF support for backbone services in 1995 eliminated the few remaining restraints on the commercialization of the Internet. The Internet's transformation from a network designed primarily to facilitate academic interchange into a medium of mass communications has made managing the Internet considerably more complicated. The Internet was once only charged with bringing together a relatively small number of fairly sophisticated, institutional users who generally shared common goals. It now must mediate among an increasingly disorderly onslaught of private users each pursuing ever more divergent objectives. This has greatly complicated traffic management, as the variability in usage patterns has increased and the beneficial effects of shared institutional norms and relationships have dwindled. This shift has also created pressure to simplify the demands imposed on end users by incorporating more of those functions into the core network.

2. The Emergence of Network-Intensive Applications

By contemporary standards, early Internet applications, such as e-mail, web access, newsgroups, and file transfer, placed fairly modest demands on the network. Overall file sizes were relatively small, and delays of a second or two typically went unnoticed. The commercialization of the Internet has spurred the development of applications which place greater demands on network services. Bandwidth-hungry applications, such as music downloads, on-line gaming, and streaming video, are placing increasing pressure on network capacity, as has the growth in telecommuting and home networking. Equally important is the emergence of applications that are less tolerant of variations in throughput rates, such as streaming media and Internet telephony.

[18] The discussion that follows draws in part on the analysis offered by Marjory S. Blumenthal & David D. Clark, *Rethinking the Design of the Internet: The End-to-End Arguments vs. the Brave New World*, 1 ACM TRANSACTIONS ON INTERNET TECH. 70 (2001), *reprinted in* COMMUNICATIONS POLICY IN TRANSITION: THE INTERNET AND BEYOND 91 (Benjamin M. Compaine & Shane Greenstein eds., 2001); *see also* Hans Kruse, William Yurcik & Lawrence Lessig, The Inter*NAT*: Policy Implications for Internet Architecture Debate (unpublished manuscript presented at the 28th Annual Telecommunications Policy Research Conference), *available at* http://www.tprc.org/abstracts00/internatpap.pdf.

These concerns have led many network providers to make the terms of interconnection vary with bandwidth usage. For example, many last-mile providers either forbid end users to use bandwidth-intensive applications, such as music downloads, streaming media, and website hosting, or instead require that they pay higher charges before doing so.[19] Similarly, backbone providers often base the amounts they charge for interconnection on volume-related considerations. Backbones who exchange traffic of roughly equal value enter into "peering" arrangements that are similar to telecommunications arrangements known as "bill and keep." Under peering arrangements, the originating backbone collects and retains all of the compensation for the transaction notwithstanding the fact that other backbones also incur costs to terminate the transaction. So long as the traffic initiated and terminated by each backbone is roughly equal in value, peering allows backbones to forego the costs of metering and billing these termination costs without suffering any adverse economic impact. Peering is not economical, however, in cases where the value of the traffic being terminated is not reciprocal. As a result, smaller-volume backbones are often required to enter into "transit" arrangements in which they must pay larger backbones compensation for terminating their traffic.[20]

The growing importance of time-sensitive applications is also placing pressure on system designers to employ "policy-based routers," which can discriminate among packets and assign them different levels of priority, depending upon the source of the packet or the nature of the application being run. This represents a marked departure from TCP/IP, which manages packets on a "first come, first served" basis and in which packets are routed without regard to the nature of the communications being transmitted.

3. The Growth in Distrust of Other Endpoints

As noted earlier, TCP/IP, which still represents the dominant suite of protocols employed by the Internet, dictates that packets be routed without regard to their source. The anonymity of this system of transmission was implicitly built on the presumption that the other endpoints in the system were relatively trustworthy and were cooperating in order to achieve common goals.

The rise of e-commerce has created the need for increased levels of confidence in the identity of the person on the other end of the connection. At the same time, end users have become increasingly

[19] Wu, *supra* note 5, at 152-54, 157-62.

[20] *See* Kende, *supra* note 12, at 47-52 (providing an overview of backbone "peering" and "transit").

frustrated by intrusions thrust upon them by other end users. Although some examples, such as spam, are relatively innocuous, other examples are considerably more malicious, such as viruses, worms, Trojan horses,[21] pornographic websites masquerading as less objectionable content, and programs that mine cookies for private information. Although end users are in a position to undertake certain measures to protect themselves against these harms, some Internet providers are interposing elements into the body of their network to shield end users from such dangers.

4. The Needs of Law Enforcement

The demands of law enforcement represent another factor that is driving the Internet away from the anonymous, fully interoperable architecture that existed in the narrowband era. For example, the Communications Assistance for Law Enforcement Act (CALEA) requires that all telecommunications carriers configure their networks in a way that permits law enforcement officials to place wiretaps on telephone calls.[22] Emerging Internet telephone systems are not easily rendered wiretap compatible. In contrast to the architecture of conventional telephone networks, which requires that all voice traffic pass through a discrete number of network gateways, Internet telephony technologies rely upon the decentralized structure inherent in the Internet. Furthermore, even if law enforcement officials found an appropriate location to intercept Internet telephone traffic, the packet anonymity inherent in TCP/IP would make it extremely difficult for law enforcement officials to separate the telephony-related packets from the other packets in the data stream. As a result, the FCC recently issued a Notice of Proposed Rulemaking and Declaratory Ruling tentatively concluding that CALEA applies to all facilities-based providers of any type of broadband Internet access service and to managed or mediated Internet telephony services.[23] Similarly, states' desire to impose sales taxes on Internet transactions may prompt them to push for changes to the architecture of the Internet to permit them to conduct some degree of monitoring of on-line commercial activity. Any solution to either problem would almost certainly require a deviation from the content and application transparency inherent in TCP/IP.

[21] Trojan horses are malicious pieces of code concealed within programs that perform beneficial functions.

[22] 47 U.S.C. § 1002(a) (2000).

[23] Communications Assistance for Law Enforcement Act and Broadband Access Services, *Notice of Proposed Rulemaking and Declaratory Ruling*, FCC 04-187, slip op. at 18-35 ¶¶ 17-59 (F.C.C. Aug. 4, 2004), *available at* http://hraunfoss.fcc.gov/edocs_public/attachmatch/FCC-04-187A1.pdf.

D. Network Neutrality Proposals

Together these changes are placing increasing pressure on last-mile broadband providers to configure their networks in ways that differentiate among packets on the basis of the source, application, or content associated with it. These moves towards discriminatory treatment have raised the concern that some providers will use their control over the last mile to harm competition. Advocates of network neutrality have advanced two different types of regulatory proposals to curb the dangers that they perceive.[24] The first, known as "multiple ISP access," would require last-mile providers to serve all ISPs on a nondiscriminatory basis. The second, sometimes called "connectivity principles," would limit last-mile providers' ability to restrict end users' ability from attaching devices, running applications, and accessing content as they see fit.

1. Multiple ISP Access

The fact that some last-mile broadband providers require their customers to connect to the Internet through their own proprietary ISP has prompted calls for the FCC to prohibit such exclusivity arrangements and to require that last-mile providers make their networks accessible to all unaffiliated ISPs on a nondiscriminatory basis. The concern is that allowing the broadband provider to control the market for ISP services has the potential to reduce consumer choice and harm competition. The opposing sides each attempted to gain the rhetorical high ground by employing terminology designed to color the way the FCC viewed the issue. Network neutrality advocates attempted to frame the issue as focusing on "open access," while broadband network owners referred to the issue as "forced access."[25] In an apparent attempt to sidestep the political overtones associated with either designation, the FCC has since framed the issue as "multiple ISP access."[26]

The FCC has vacillated on multiple ISP access over the course of various merger clearance proceedings.[27] The agency initially rejected

[24] *See* Philip J. Weiser, *Toward a Next Generation Regulatory Strategy*, 35 LOY. U. CHI. L.J. 41, 44-48 (2004) (distinguishing between the two approaches to network neutrality); Wu, *supra* note 5, at 147-50 (same).

[25] *See* Applications for Consent to Transfer of Control of Licenses and Section 214 Authorizations from MediaOne Group, Inc., Transferor, to AT&T Corp., Transferee, *Memorandum Opinion and Order*, 15 FCC Rcd. 9816, 9866 ¶ 114 (2000) [hereinafter *AT&T-MediaOne Merger*].

[26] *See Cable Modem Declaratory Ruling and NPRM*, *supra* note 2, at 4839 ¶ 72.

[27] For a more detailed review of the regulatory history of multiple ISP access, see Christopher S. Yoo, *Vertical Integration and Media Regulation in the New Economy*, 19 YALE J. ON REG. 171, 251-52 (2002); Spulber & Yoo, *supra* note 16, at 1015-18.

calls for multiple ISP access when clearing AT&T's acquisitions of TCI and MediaOne,[28] only to backtrack somewhat by acceding to a multiple ISP access requirement imposed by the Federal Trade Commission (FTC) during the America Online-Time Warner merger.[29] Since then, the FCC has returned to its original position, declining to impose multiple ISP access when approving the sale of AT&T's cable properties to Comcast.[30] At the same time, the FCC has successfully forestalled attempts by cities to impose multiple ISP access either as a matter of municipal ordinances[31] or as part of their approval of the transfer of licenses needed to complete these mergers[32] on the grounds that such regulation falls within the exclusive jurisdiction of the federal government. Throughout these preemption disputes, the FCC continued to emphasize that it had not yet determined whether to impose open access and asked the courts not to resolve the issue.[33]

It is only recently that the FCC has finally begun to address the issue in earnest. In the ongoing cable modem proceedings, the FCC has twice requested comment on the advisability of requiring cable modem systems to provide multiple ISP access.[34] It also raised the issue in the ongoing wireline broadband proceedings, seeking comment on whether it should impose multiple ISP access on DSL providers in the event that it decided to exempt them from the unbundled network element (UNE)

[28] *AT&T-MediaOne Merger*, *supra* note 25, at 9866 ¶¶ 114-115; Applications for Consent to Transfer of Control of Licenses and Section 214 Authorizations from Tele-Communications, Inc., Transferor, to AT&T Corp., Transferee, *Memorandum Opinion and Order*, 14 FCC Rcd. 3160, 3197-98 ¶ 75 (1999) [hereinafter *TCI-AT&T Merger*].

[29] Applications for Consent to Transfer of Control of Licenses and Section 214 Authorizations by Time Warner, Inc. and America Online, Inc., Transferors, to AOL Time Warner Inc., Transferee, *Memorandum Opinion and Order*, 16 FCC Rcd. 6547, 6568-69 ¶¶ 57-58 (2001) [hereinafter *Time Warner-AOL Merger*]; America Online, Inc., *Decision & Order,* No. C-3989, slip op. at 2, 6-9, 11-17 (F.T.C. Dec. 18, 2000), *available at* http://www.ftc.gov/os/2000/12/aoldando.pdf.

[30] Applications for Consent to Transfer of Control of Licenses from Comcast Corp. and AT&T Corp., Transferors, to AT&T Comcast Corp., *Memorandum Opinion and Order*, 17 FCC Rcd. 23,246, 23,300-01 ¶ 135 (2002).

[31] *See* Comcast Cablevision of Broward County, Inc. v. Broward County, 124 F. Supp. 2d 685, 686-87 (S.D. Fla. 2000).

[32] *See* MediaOne Group, Inc. v. County of Henrico, 257 F.3d 356, 360 (4th Cir. 2001); AT&T Corp. v. City of Portland, 216 F.3d 871, 875 (9th Cir. 2000).

[33] *See* Amicus Curiae Brief of the Federal Communications Commission at 15-18, *MediaOne Group, Inc. v. County of Henrico* (Nos. 00-1680(L), 00-1709, 00-1719) (available at 2000 WL 33991834); Brief of the Federal Communications Commission as Amicus Curiae at 19-26, 30-31, *AT&T Corp. v. City of Portland* (No. 99-35609) (available at 1999 WL 33631595).

[34] The FCC made its initial request in 2000 when issuing its Notice of Inquiry in the cable modem proceeding. *See Cable Modem NOI*, *supra* note 2, at 19,298-306 ¶¶ 25-49. It reiterated the request in its subsequent Declaratory Ruling and Notice of Proposed Rulemaking in 2002. *See Cable Modem Declaratory Ruling and NPRM*, *supra* note 2, at 4839-41 ¶¶ 72-74, 4843-47 ¶¶ 80-93.

access requirements imposed by the Telecommunications Act of 1996.[35] The FCC's request for comments would prove prescient, as the subsequent Triennial Review Order would eventually strike most DSL-related facilities from the list of network elements to which telecommunications carriers have the right of unbundled access.[36]

A number of entities have submitted comments calling upon the FCC to mandate multiple ISP access.[37] An alliance of trade associations representing the computer, telecommunications equipment, semiconductor, consumer electronics, software and manufacturing sectors known as the High Tech Broadband Coalition (HTBC)[38] has offered a more limited proposal, which calls for the FCC to require DSL providers to honor any existing access agreements with unaffiliated ISPs and to make any arrangements with their affiliated ISPs available to unaffiliated ISPs in a nondiscriminatory manner for a period of at least two years.[39]

2. Connectivity Principles

Other proposals have shifted their attention away from preserving ISP competition and have instead focused on preserving competition among content and applications providers. For example, Professors Timothy Wu and Lawrence Lessig have proposed a network neutrality regime that would prohibit last-mile providers from imposing any restrictions on end users' ability to run the applications, attach the devices, and access the content of their own choosing except those restrictions that are necessary to comply with a legal duty, prevent

[35] *See Wireline Modem NPRM*, *supra* note 2, at 3042-43 ¶¶ 50-52.

[36] Competitors remain free, however, to obtain unbundled access to the entire local loop and provide both voice and DSL services. *See* Review of the Section 251 Unbundling Obligations of Incumbent Local Exchange Carriers, *Report & Order & Order on Remand & Further Notice of Proposed Rulemaking*, 18 FCC Rcd. 16,978, 17,131-36 ¶¶ 255-63 (2003), *aff'd in relevant part sub nom.* U.S. Telecom Ass'n v. FCC, 359 F.3d 554, 578-85 (D.C. Cir. 2004).

[37] *See* Comments of Amazon.com at 9, *Cable Modem Declaratory Ruling and NPRM* (F.C.C. filed June 17, 2002) (CS Dkt. No. 02-52), *available at* http://gullfoss2.fcc.gov/prod/ecfs/retrieve.cgi?native_or_pdf=pdf&id_document=6513198055; EarthLink, Inc. Comments in CS Docket No. 02-52 at 3-14, *Cable Modem Declaratory Ruling and NPRM* (F.C.C. filed June 17, 2002) (CS Dkt. No. 02-52), *available at* http://gullfoss2.fcc.gov/prod/ecfs/retrieve.cgi?native_or_pdf=pdf&id_document=6513198478.

[38] The specific trade associations include the Business Software Alliance, Consumer Electronics Association, Information Technology Industry Council, National Association of Manufacturers, Semiconductor Industry Association, and Telecommunications Industry Association. It has the active support of such companies as Intel, Alcatel, Catera, and Corning.

[39] Reply Comments of High Tech Broadband Coalition at i-ii, 6-8, Appropriate Framework for Broadband Access to the Internet over Wireline Facilities (F.C.C. filed July 1, 2002) (CC Dkt. No. 02-33).

physical harm to the network, prevent interference with other users' connections, ensure quality of service, and prevent violations of security.[40]

HTBC has advanced a similar proposal that would impose a series of "connectivity principles" on all last-mile broadband providers. This proposal would require that all last-mile broadband providers give end users unrestricted access to all content and allow them to run any applications and attach any devices they desire, so long as these efforts do not harm the providers' networks, enable theft of services, or exceed the bandwidth limitations of the particular service plan.[41] The HTBC's proposal has drawn the support of a group composed primarily of software and content providers known as the Coalition of Broadband Users and Innovators (CBUI).[42] FCC Chairman Michael Powell sounded similar themes when called upon the broadband industry to embrace a series of "Internet Freedoms." In sharp contrast to the HTBC's proposal, however, Powell's vision would arise through voluntary conduct rather than through regulation.[43]

II. UNDERSTANDING THE ECONOMICS OF END-TO-END

As noted earlier, network neutrality advocates have drawn much of the inspiration for their regulatory proposals from the end-to-end argument pioneered by Saltzer, Reed, and Clark. Simply put, the end-to-end argument counsels against introducing intelligence into the core of the Internet and in favor of restricting higher levels of functionality to the servers operating at the edges of the network. The "pipes" that constitute the core of the network should be kept "dumb" and should focus solely on passing along packets as quickly as possible. Part A describes the basic intuitions underlying the end-to-end argument. Part B undertakes a close analysis of the implications of the end-to-end argument for the major regulatory proposals, concluding that network neutrality proposals are based on an over reading of Saltzer, Reed, and

[40] *Ex parte* Letter of Timothy Wu and Lawrence Lessig, *Cable Modem Declaratory Ruling and NPRM*, at 12-15 (F.C.C. filed Aug. 22, 2003) (CS Docket No. 02-52), *available at* http://gullfoss2.fcc.gov/prod/ecfs/retrieve.cgi?native_or_pdf=pdf&id_document=6514683884; *see also* Wu, *supra* note 5, at 165-72.

[41] Comments of the High Tech Broadband Coalition at 6-9, *Cable Modem Declaratory Ruling & NPRM* (F.C.C. filed June 17, 2002) (CC Dkt. No. 02-52), *available at* http:// gullfoss2.fcc.gov/prod/ecfs/retrieve.cgi?native_or_pdf=pdf&id_document=6513198026.

[42] *Ex parte* Communication from the Coalition of Broadband Users and Innovators at 3-4, *Cable Modem Declaratory Ruling and NPRM* (F.C.C. filed Jan. 8, 2003) (CS Dkt. No. 02-52), *available at* http://gullfoss2.fcc.gov/prod/ecfs/retrieve.cgi?native_or_pdf=pdf&id_document=6513401671. CBUI includes such notable content and software providers as Microsoft, Disney, Amazon.com, eBay, and Yahoo!, as well as the Media Access Project, the Consumer Electronics Association, and the National Association of Manufacturers.

[43] *See generally* Michael K. Powell, *The Digital Broadband Migration: Toward a Regulatory Regime for the Internet Age*, 3 J. ON TELECOMM. & HIGH TECH. L. 5 (2004).

Clark's work that expands it far outside its proper scope. In fact, a careful examination of the rationale underlying the end-to-end argument reveals that it is fundamentally incompatible with network neutrality advocates' attempts to turn the end-to-end argument into a regulatory mandate.

A. The Classic Statement of the End-to-End Architecture

The fundamental logic of the end-to-end argument is most easily understood by examining the core illustration offered by Saltzer, Reed, and Clark to articulate it: careful file transfer, in which a file stored on the hard drive of computer *A* is transferred to the hard drive of computer *B* without errors.[44] Roughly speaking, this function can be divided into five steps:

1. Computer *A* reads the file from its hard disk and passes it to the file transfer program.
2. The file transfer program running on computer *A* prepares the file for transmission by dividing it into packets and hands off the packets to the data communication network.
3. The data communication network moves the packets from computer *A* to computer *B*.
4. The file transfer program running on computer *B* reassembles the packets into a coherent file.
5. The file transfer program saves the file onto computer *B*'s hard disk.

Errors can emerge at any step in this process. Computer *A* can misread the file from the hard disk. The file transfer program on Computer *A* can introduce mistakes when copying the data from the file. The communication network can drop or change bits in a packet or lose a packet altogether. The file transfer program on Computer *B* can also produce errors when converting the packets back into a coherent file. Computer *B* can miswrite the file to its hard disk. The transfer can also be jeopardized by larger-scale hardware or software failures.

Saltzer, Reed, and Clark compare two different approaches to managing the risk of such errors. One approach is to perform error checking at each intermediate step along the way. The other approach is known as "end-to-end check and retry." Under this approach, no error

[44] The discussion that follows is based on Saltzer et al., *supra* note 4, at 278-80.

checking is performed at any of the intermediate steps. Instead, the only error checking occurs when the terminating end of the process (computer *B*) verifies the accuracy of the file transfer with the initiating end (computer *A*) after the entire transaction has been completed.

The core conclusion of Saltzer, Reed, and Clark's work is that system designers should adopt a presumption in favor of the latter approach. They base their argument on two insights. First, no matter how many intermediate error checks are introduced, the terminating end of the file transfer must still verify the transaction with the originating end after all of the steps have been completed. The fact that such end-to-end verification is necessary no matter what other intermediate reliability measures are built into the system renders any additional measures redundant, thus raising doubts as to the justifiability of any additional measures.[45]

Second, intermediate error checking should properly be regarded as an engineering tradeoff between reliability and performance. Errors can be reduced, but only at the cost of introducing a degree of redundancy into the network that will have the inevitable effect of slowing it down. Saltzer, Reed, and Clark emphasize that different applications vary in their tolerance for unreliability as well as their demand for speed. Imposing reliability checks in low-level subsystems that are common to all applications may have the uneconomical result of forcing all applications to incur the performance costs even if the increase in reliability does not provide particular applications with commensurate benefits.[46]

Together these insights suggest that system designers should avoid designing higher-level functions into routers located in the core of the network. Instead, the Internet should presumptively be engineered with any such functions concentrated in the servers that operate at the network's edge. Saltzer, Reed, and Clark extend the same basic rationale to other system functions, such as delivery guarantees, secure transmission of data, duplicate message suppression, and transaction management.[47]

B. End-to-End as a Case-by-Case Approach

Network neutrality proponents contend that the end-to-end argument justifies prohibiting Internet providers from introducing additional degrees of intelligence into their core networks. In short, all of the intelligence should be restricted to the servers operating at the

[45] *Id.* at 281.

[46] *Id.*

[47] *Id.* at 282-84.

edge of the network. They also argue that the end-to-end argument mandates that all broadband network owners employ protocols like TCP/IP that ensure that the core of the network remains relatively transparent and dumb.[48]

Although the end-to-end argument does support a presumption against introducing higher-level functions into the network's core, it does not justify elevating this presumption into an inviolable precept. Conceding that it is "too simplistic to conclude that the lower levels should play no part in obtaining reliability,"[49] Saltzer, Reed, and Clark's original article articulating the end-to-end argument squarely concludes that "the end-to-end argument is not an absolute rule, but rather a guideline that helps in application and protocol design analysis."[50] In fact, the cost-performance tradeoff underlying the end-to-end argument requires "subtlety of analysis" and can be "quite complex."[51] Indeed, a later article by the same authors responding to calls for allowing the core of the Internet to exercise a greater level of functionality explicitly recognizes that "[t]here are some situations where applying an end-to-end argument is counterproductive"[52] and concludes that the proper approach is to "take it case-by-case."[53] The end-to-end argument is thus more properly regarded as merely "one of several important organizing principles for systems design" rather than as an absolute.[54] Although Saltzer, Reed, and Clark suggest that deviations from it will be rare, they acknowledge that "there will be situations where other principles or goals have greater weight."[55]

Other technologists have drawn similar conclusions. One of the original authors of the end-to-end argument, writing with Marjory Blumenthal, candidly acknowledges that "the end-to-end arguments are not offered as an absolute" and that "[t]here are functions that can only be implemented in the core of the network."[56] Indeed, they argue that the developments described in Section I have made the case for introducing greater intelligence into Internet's core all the more

[48] *See, e.g.*, Lemley & Lessig, *supra* note 5, at 931-32.

[49] Saltzer et al., *supra* note 4, at 280.

[50] *Id.* at 285.

[51] *Id.* at 284. To take but one example, the desirability of end-to-end depends in part on the length of the file. If a system drops one message per one hundred messages sent, the probability that all packets will arrive correctly decreases exponentially as the length of the file increases (and thus the number of packets composing the file) increases. *Id.* at 280-81.

[52] David P. Reed et al., *Commentaries on "Active Networking and End-to-End Arguments,"* 12 IEEE NETWORK 69, 69 n.1 (1998).

[53] *Id.* at 70.

[54] *Id.*

[55] *Id.*

[56] Blumenthal & Clark, *supra* note 18, at 71.

compelling. They conclude, apparently with the concurrence of Saltzer,[57] that in many cases "an end-to-end argument isn't appropriate in the first place."[58] Samrat Bhattacharjee, Kenneth Calvert, and Ellen Zegura conclude that the end-to-end argument "do[es] *not* rule out support for higher-level functionality within the networks" and instead simply requires that the costs and benefits inherent in the engineering tradeoff be carefully evaluated.[59] Indeed, there are services that depend on information that is only available inside the network and thus cannot exist without relying to some degree on what has been called "active networking."[60] Dale Hatfield acknowledges that the desire to improve the security, manageability, scalability, and reliability of the Internet may justify introducing greater intelligence into the core of the network.[61] As a result, Hatfield argues against allowing regulation that prevents network owners from deviating from the end-to-end architecture and instead simply warns that deviations from the end-to-end argument should be undertaken with extreme care.[62]

At this point, the incongruity of invoking the end-to-end argument as support for network neutrality as a regulatory mandate should be apparent. Far from justifying an absolute prohibition against placing intelligence in the core of the network, the end-to-end argument stands squarely opposed to such a simplistic approach.[63] Simply put, a close analysis of the end-to-end argument reveals that it does not support the proposition for which many network neutrality proponents invoke it. Indeed, as Marjory Blumenthal has noted, this incongruity demonstrates the extent to which network neutrality advocates' embrace of the end-to-end argument has left the realm of cost-benefit analysis and has instead

[57] *See id.* at 102 n.19 (citing personal communication with Jerome Saltzer as support for this proposition).

[58] *Id.* at 80.

[59] Samrat Bhattacharjee et al., *Active Networking and the End-to-End Argument*, 1997 PROC. INT'L CONF. ON NETWORK PROTOCOLS 220, 221.

[60] *Id.*; *see also* Samrat Bhattacharjee et al., *Commentaries on "Active Networking and End-to-End Arguments*," 12 IEEE NETWORK 66 (1998).

[61] Dale N. Hatfield, *Preface*, 8 COMMLAW CONSPECTUS 1, 3 (2000).

[62] *Id.*

[63] Although the end-to-end argument only supports a case-by-case approach to network design, it is arguable that such cases will prove so rare that the costs of evaluating the merits of each individual case exceed the benefits of doing so. Such categorical balancing is particularly perilous in industries, such as broadband, that are in a state of technological and economic flux. Even if regulators were to strike the proper balance today, the underlying technological and economic context would soon shift. A real danger exists that this inherent lag will cause regulation intended to promote economic efficiency to inhibit it. *See, e.g.*, STEPHEN BREYER, REGULATION AND ITS REFORM 286-87 (1982); 2 ALFRED E. KAHN, THE ECONOMICS OF REGULATION 127 (1971); Richard A. Posner, *Natural Monopoly and Its Regulation*, 21 STAN. L. REV. 548, 611-15 (1969). Such concerns counsel strongly in favor of allowing private ordering rather than the government to determine network configurations.

entered the realm of ideology.[64] As a result, it is critical that network neutrality proposals not evade critical analysis by masquerading as nothing more than the application of sound engineering principles.

The foregoing discussion casts a new and somewhat ironic light on Lessig's observation that "code is law."[65] The point Lessig was attempting to make was that the architecture enshrined in the Internet's communications protocols can have as dramatic an impact on competition and innovation as direct regulation. Most network neutrality advocates have failed to appreciate that this admonition cuts both ways.[66] While it is true that allowing Internet providers to impose proprietary protocols could have a significant impact on innovation and competition, forbidding them from doing so could have equally dramatic effects. Either decision necessarily involves policymakers in the unenviable task of picking technological winners and losers. The impossibility of technologically neutral government intervention undercuts claims that imposing the end-to-end argument as a regulatory mandate represents the proper way to show humility about the future of the Internet.[67]

Not only does government-imposed network neutrality contradict the letter of the end-to-end argument, it turns Lessig's admonition on its head. Lessig intended the statement to indicate how the architecture of the Internet could constitute a private substitute for many of the functions previously served by law. Indeed, Lessig warned of the dangers of allowing the government to dictate the standards that must be included in Internet code.[68] It would be a strange inversion of this argument to give the phrase "code is law" literal rather than figurative meaning and to sanction greater governmental control over the architecture of the Internet.

III. THE INTERRELATIONSHIP BETWEEN NETWORK NEUTRALITY AND THE ECONOMICS OF VERTICAL INTEGRATION

In addition to misunderstanding the proper scope of the end-to-end argument, network neutrality proponents have largely overlooked the close relationship between their proposals and the economics of vertical integration. This section examines how vertical integration theory sheds new light on the debates surrounding network neutrality. Part A reviews

[64] Marjory S. Blumenthal, *End-to-End and Subsequent Paradigms*, 2002 L. REV. MICH. ST. U. DET. C.L. 709, 710 (2002).

[65] LAWRENCE LESSIG, CODE AND OTHER LAWS OF CYBERSPACE 6 (1999).

[66] For a notable exception, see Wu, *supra* note 5, at 148.

[67] *See* LESSIG, *supra* note 5, at 35, 39.

[68] *See* Lawrence Lessig, *The Limits in Open Code: Regulatory Standards and the Future of the Net*, 14 BERKELEY TECH. L.J. 759, 764-67 (1999).

the structure of the broadband industry and describes how network neutrality is designed to redress the supposed problems caused by vertical integration. The relationship between network neutrality and vertical integration is clear whether one conceives of the broadband industry as consisting of a traditional, three-step chain of production implicit in multiple ISP access proposals or whether one follows the more recent trend of describing the broadband industry as consisting of a series of horizontal "layers" underlying the regulatory approach embodied in the connectivity principles.

Part B reviews the key insights of vertical integration theory. It is now widely recognized that vertical integration can create economic harms only if certain structural preconditions are met. An empirical analysis reveals that these structural preconditions are not satisfied with respect to the broadband industry. This in turn undermines claims that the types of vertical integration that network neutrality is designed to foreclose poses a serious policy concern.

A. Two Conceptions on the Structure of the Broadband Industry

The major network neutrality proposals have embedded within them two, rather different conceptions of the vertical structure of the broadband industry. Multiple ISP access proposals implicitly conceive of providers being organized in a traditional, three-step chain of distribution, in which the ISPs act as a wholesaler and the last-mile providers play the role of the retailer. The proponents of connectivity principles conceive of the broadband industry as consisting of a series of layers.

1. The Conventional Vertical Market Structure Implicit in Multiple ISP Access

Although the structure of the broadband industry may at times seem mysterious, it is in fact quite ordinary when viewed from a certain perspective.[69] Its basic organization differs little from that of the typical manufacturing industry, which is divided into a three-stage chain of production. The first and last stages are easiest to understand. The manufacturing stage is occupied by companies that create the actual products to be sold. The retail stage consists of those companies responsible for the final delivery of the products to end-users. Although it is theoretically possible for retailers to purchase products directly from manufacturers, in practice logistical complications often give rise to an intermediate stage mediating between manufacturers and retailers.

[69] The following discussion is adapted from Yoo, *supra* note 27, at 182, 250-51.

Firms operating in this intermediate stage, known as wholesalers, purchase goods directly from manufacturers and assemble them into complete product lines and distribute them to retailers.

Despite claims that the Internet is fundamentally different from other media, the broadband industry mapped comfortably onto this three-stage vertical market structure. The manufacturing stage encompasses those companies that generate the webpage content and Internet-based services that end users actually consume. The wholesale stage is occupied by the ISPs and backbone providers, which aggregate content and applications. Finally, last-mile providers deliver the content and service packages assembled by the ISPs to end customers.

The proponents of multiple ISP access in essence are concerned that vertical integration between the retail and wholesale levels of this chain of distribution will allow network owners to use the leverage provided by their control of the retail stage to harm competition at the wholesale level. In other words, they argue that allowing last-mile providers to deny unaffiliated ISPs access to their customers threatens ISP competition.[70]

2. The "Layered" Approach Implicit in Connectivity Principles

The connectivity principles implicitly rely on what has become known as the "layered model" to Internet regulation.[71] This approach disaggregates networks into four horizontal layers that cut across different network providers.[72] The bottommost layer is the *physical*

[70] *See, e.g.*, Hausman et al., *supra* note 7, at 158-65; Lemley & Lessig, *supra* note 5, at 940-43; Rubinfeld & Singer, *supra* note 5, at 664-70.

[71] *See* Kevin Werbach, *A Layered Model for Internet Privacy*, 1 J. ON TELECOMM. & HIGH TECH. L. 37, 57-64 (2002); Richard Whitt, *A Horizontal Leap Forward: Formulating a New Communications Public Policy Framework Based on the Network Layers Model*, 56 FED. COMM. L.J. 587, 624 (2004). For other leading discussions analyzing the Internet through the layered model, see LESSIG, *supra* note 5, at 23-25; Yochai Benkler, *From Consumers to Users: Shifting the Deeper Structures of Regulation Towards Sustainable Commons and User* Access, 52 FED. COMM. L.J. 561 (2000); Douglas C. Sicker & Joshua L. Mindel, *Refinements of a Layered Model for Telecommunications Policy*, 1 J. ON TELECOMM. & HIGH TECH. L. 69 (2002); Timothy Wu, *Application-Centered Internet Analysis*, 85 VA. L. REV. 1163, 1189-92 (1999). For a different vision of layered competition, see Timothy F. Bresnahan, *New Modes of Competition: Implications for the Future Structure of the Computer Industry*, *in* COMPETITION, INNOVATION AND THE MICROSOFT MONOPOLY: ANTITRUST IN THE DIGITAL MARKETPLACE 155 (Jeffrey A. Eisenach & Thomas M. Lenard eds., 1999).

[72] The layered model is related to the Open Systems Interconnection (OSI) model developed by the International Standards Organization (ISO) in the 1980s, which divides seven different layers: application, presentation, session, transport, network, data link, and physical. Some of these distinctions between those layers have greater relevance for technologists than for policy analysts. *See* Werbach, *supra* note 71, at 59.

FIGURE 3
THE LAYERED MODEL OF BROADBAND ARCHITECTURE

CONTENT LAYER (*e.g.*, individual e-mail, webpages, voice calls, video programs)
APPLICATIONS LAYER (*e.g.*, web browsing, e-mail, Internet telephony, streaming media, database services)
LOGICAL LAYER (*e.g.*, TCP/IP, domain name system, telephone number system)
PHYSICAL LAYER (*e.g.*, telephone lines, coaxial cable, backbones, routers, servers)

layer, which consists of the hardware infrastructure that actually carries the communications. The second layer is the *logical layer*, which is composed of the protocols responsible for organizing the management and routing functions of the network. The third layer is the *applications layer*, comprised of the particular programs and functions used by consumers. The fourth layer is the *content layer*, which consists of the particular data being conveyed.

The distinction between the layers can be illustrated in terms of the most common Internet application: e-mail. Assuming that the particular e-mail in question is sent via DSL, the physical layer consists of the telephone lines, e-mail servers, routers, and backbone facilities needed to convey the e-mail from one location to another. The logical layer consists of the SMTP protocol employed by the network to route the e-mail to its destination. The application layer consists of the e-mail program used, such as Microsoft Outlook. The content layer consists of the particular e-mail message sent.

The connectivity principles are motivated by a concern that last-mile providers will use their control of the physical layer to reduce competition in the application and content layer by deviating from TCP/IP currently employed in the logical layer and replacing it with proprietary, noninteroperable protocols. The connectivity principles are designed to forestall this dynamic by mandating that last-mile providers adhere to nonproprietary protocols and to open their networks to all applications and content on a nondiscriminatory basis.[73]

[73] *See also id.* at 65-66 (arguing that the layered model requires that interfaces between each layer be kept open).

B. Market Structure and Vertical Integration

Vertical integration has long been a source of economic controversy.[74] Until the 1970s, competition policy generally viewed vertical integration with considerable hostility. The emergence of the Chicago School of antitrust law and economics raised serious doubts regarding the preexisting orthodoxy, arguing that a monopolist would have little to gain by vertically integrating. In addition, certain structural preconditions must be satisfied before vertical integration can harm competition. Specifically, both the upstream and downstream markets that are being brought together through vertical integration must be concentrated and protected by barriers to entry. If not, vertical integration should be permitted.

These developments in turn prompted the emergence of a post-Chicago School, which contradicted the Chicago School by identifying circumstances in which vertical integration can harm competition. While disagreeing over many key aspects of vertical integration theory, the post-Chicago School implicitly agreed that the same structural preconditions must be met before vertical integration can plausibly be problematic.[75] The fact that these structural preconditions are enshrined in the Merger Guidelines promulgated by the Justice Department and the FTC demonstrates the broad acceptance that these principles now enjoy.[76]

Applying these principles to the broadband industry strongly suggests that the FCC should not erect what would amount to a per se bar to vertical integration. Considering first the requirement that the primary market be concentrated, the Merger Guidelines employs a measure of concentration known as the Hirschman-Herfindahl index (HHI) that has become the standard concentration under modern competition policy. HHI is calculated by adding the square of the market share of each competitor.[77] The result is a continuum that places

[74] The discussion that follows is based on the more complete presentation at Yoo, *supra* note 27, at 253-68. For a review of the historical development of vertical integration theory presented, see *id.* at 185-205.

[75] Specifically, post-Chicago scholarship typically models the relevant markets either as dominant firm industries or as oligopolies engaged in Cournot or Bertrand competition. Both of these approaches require that the relevant markets be highly concentrated and protected by barriers to entry. Yoo, *supra* note 27, at 203-05, 265-67.

[76] *See* U.S. DEPARTMENT OF JUSTICE & FEDERAL TRADE COMMISSION, *Non-Horizontal Merger Guidelines, in* 1992 HORIZONTAL MERGER GUIDELINES, §§ 4.131, 4.212, 57 Fed. Reg. 41,552 (1992), *available at* http://www.usdoj.gov/atr/public/guidelines/2614.htm [hereinafter *Non-Horizontal Merger Guidelines*] (requiring that the relevant markets be concentrated); *id.* §§ 4.132, 4.133, 4.21 (requiring that the relevant markets be protected by barriers to entry).

[77] For example, a market of four firms with market shares of 30%, 30%, 20% and 20%, respectively, would have an HHI of $30^2 + 30^2 + 20^2 + 20^2 = 900 + 900 + 400 + 400 = 2600$.

the level of concentration on a scale from 0 (in the case of complete market deconcentration) to 10,000 (in the case of monopoly). The Guidelines indicate that the antitrust authorities are unlikely to challenge a vertical merger unless HHI in the primary market exceeds 1800, which is the level of concentration that would result in a market comprised of between five and six competitors of equal size.[78]

Determining whether the market is concentrated depends on a proper market definition, which in turn requires the identification of the relevant product and geographic markets.[79] Defining the relevant product market is relatively straightforward: empirical evidence indicates that broadband represents an independent product market that is distinct from narrowband services.[80] Defining the relevant geographic market has proven more problematic. Many analyses have mistakenly assumed that the relevant geographic market is the local market in which last-mile broadband providers meet end users. Because these markets are typically dominated by two players—the incumbent cable operators selling cable modem service and the incumbent local telephone company offering DSL service—defining the geographic market in this manner yields HHIs well in excess of 4000.[81]

The problem with this analysis is that network neutrality proposals are designed to limit the exercise of market power not in the final downstream market in which last-mile providers meet end users, but rather in the upstream market in which last-mile providers meet ISPs and content/application providers. The following thought experiment

[78] *Non-Horizontal Merger Guidelines*, *supra* note 76, §§ 4.131, 213. Note that the relevant threshold for vertical mergers is more lenient than the HHI thresholds applicable to horizontal mergers. Under the Horizontal Merger Guidelines, markets with HHIs between 1000 and 1800 are regarded as "moderately concentrated" and thus "potentially raise significant competitive concerns." U.S. DEPARTMENT OF JUSTICE & FEDERAL TRADE COMMISSION, 1992 HORIZONTAL MERGER GUIDELINES § 1.51(b), 57 Fed. Reg. 41,552 (1992), *revised*, 4 Trade Reg. Rep. (CCH) ¶ 13,104 (Apr. 8, 1997), *available at* http://www.usdoj.gov/atr/public/guidelines/horiz_book/hmg1.html [hereinafter HORIZONTAL MERGER GUIDELINES]. Because vertical mergers are less likely than horizontal mergers to harm competition, the Merger Guidelines apply a more lenient HHI threshold to vertical integration. *Non-Horizontal Merger Guidelines*, *supra* note 76, § 4.0. The Merger Guidelines also reserve the possibility of challenging a vertical merger at HHI levels below 1800 if "effective collusion is particularly likely." *Id.* § 4.213. Such problems are more properly regarded as horizontal rather than vertical in nature.

[79] HORIZONTAL MERGER GUIDELINES, *supra* note 78, §§ 1.0-1.3.

[80] *See Time Warner-AOL Merger*, *supra* note 29, at 78-88 ¶¶ 69-73; Jerry A. Hausman et al., *Cable Modems and DSL: Broadband Internet Access for Residential Customers*, 91 AM. ECON. REV. 302, 303-04 (2001).

[81] *See* Amendment of Parts 1, 21, 73, 74 & 101 of the Commission's Rules to Facilitate the Provision of Fixed & Mobile Broadband Access, Educational & Other Advanced Servs. in the 2150-2162 & 2500-2690 Mhz Bands, *Notice of Proposed Rule Making & Memorandum Opinion & Order*, 18 FCC Rcd. 6722, 6774-75 ¶¶ 123-124 (2003); Hausman et al., *supra* note 7, at 155; Rubinfeld & Singer, *supra* note 5, at 649.

confirms this insight: Suppose that every last-mile provider were required to sell their proprietary interests in ISPs, application providers, and content providers. Such a change would not affect the economic relationship between end users and last-mile providers; end users seeking to purchase last-mile services would still face a de facto duopoly even if the broadband industry were completely vertically disintegrated. Compelled vertical disintegration would, however, substantially change the bargaining power between last-mile providers and ISPs and content/application providers.

It is thus this upstream market in which last-mile providers meet ISPs and providers of Internet content and applications that represents the true target of network neutrality proposals. This market is properly regarded as national in scope.[82] Major web-based providers, such as Amazon.com or eBay, are focused more on the total customers they are able to reach nationwide than they are on their ability to reach customers located in any specific metropolitan area. Their inability to reach certain customers is of no greater concern, however, than the inability of manufacturers of particular brands of cars, shoes, or other conventional goods to gain access to all parts of the country. Being cut off from certain distribution channels should not cause economic problems, so long as those manufacturers are able to obtain access to a sufficient number of customers located elsewhere. The proper question is thus not whether the broadband transport provider wields oligopoly power over broadband users in any particular city, but rather whether that provider has market power in the national market for obtaining broadband content.

When the relevant inquiry is properly framed as the national market, it becomes clear that the market is too unconcentrated for vertical integration to pose a threat to competition. The HHI is 1079, well below the 1800 threshold for vertical integration to be a source of economic concern. In addition, the two largest broadband providers (Comcast and SBC) control only 21% and 14% of the national market respectively. Absent collusion or some other impermissible horizontal practice (which would be a basis for sanction independent of concerns about vertical integration), the national broadband market is sufficiently unconcentrated to vitiate concerns about the vertical integration in the broadband industry.

In addition, the precondition that the secondary markets be concentrated and protected by entry barriers is also not met.[83] As the FCC has recognized, the market for ISPs has long been quite

[82] Yoo, *supra* note 27, at 253-54.

[83] *See id.* at 259.

FIGURE 4
LAST-MILE BROADBAND SUBSCRIBERS AS OF YEAR END 2003

Provider	Technology	Subscribers (000s)	Share	HHI
Comcast	cable modem	5,284	21%	442
SBC	DSL	3,516	14%	196
Time Warner Cable	cable modem	3,228	13%	165
Verizon	DSL	2,319	9%	85
Cox	cable modem	1,999	8%	63
Charter	cable modem	1,566	6%	39
BellSouth	DSL	1,462	6%	34
Cablevision	cable modem	1,057	4%	18
Adelphia	cable modem	960	4%	14
Qwest	DSL	637	3%	6
Bright House	cable modem	625	2%	6
Covad	DSL	517	2%	4
Sprint	DSL	304	1%	1
Mediacom	cable modem	280	1%	1
Insight	cable modem	230	1%	1
RCN	cable modem	200	1%	1
Alltel	DSL	153	1%	0
Cable One	cable modem	134	1%	0
Cincinnati Bell	DSL	99	0%	0
Century Tel	DSL	83	0%	0
Other		503	2%	1
Total		25,136	100%	1078

Source: *Fiber Faces the Inevitable Shakeout, DSL Competition*, FIBER OPTICS NEWS, Mar. 17, 2004.

competitive, and entry into ISP services has historically been quite easy.[84] Similarly, the markets for applications and content have long been the most competitive segments of the entire industry, marked by low levels of concentration and low barriers to entry. The failure to satisfy these structural preconditions renders implausible any claims that vertical integration in the broadband industry constitutes a threat to competition.

[84] *TCI-AT&T Merger*, *supra* note 28, at 3206 ¶ 93(1999).

IV. THE POTENTIAL BENEFITS OF NETWORK DIVERSITY

Conventional economic theory thus indicates that allowing last-mile providers to vertically integrate by entering into exclusivity arrangements with respect to certain content and applications providers or by requiring the use of proprietary ISPs is unlikely to harm competition. In this section, I raise a number of points that have yet to appear in either the academic literature or in the filings in the ongoing broadband proceedings before the FCC. These points show how allowing last-mile broadband providers to deviate from the principles of network neutrality can actually benefit consumers.[85] Part A examines the economic efficiencies that can result from vertical integration. Part B discusses how allowing network owners to deviate from complete interoperability can increase economic welfare by increasing the diversity of products available. Conversely, imposing network neutrality as a regulatory matter may actually have the effect of reducing innovation and limiting consumer choice by skewing the Internet towards certain types of applications and away from others. Part C analyzes the impact that connectivity principles can have on the concentration of last-mile technologies, which looms as a far more central threat to the competitive performance of the Internet than does the robustness of competition among content and applications providers. It also details how standardizing network protocols can reinforce the supply-side and demand-side economies of scale that are the primary impetus towards concentration in last-mile services. By forcing broadband providers to compete solely on price and network size, network neutrality reinforces the advantages already enjoyed by the largest players. Conversely, allowing network heterogeneity can provide new last-mile platforms, such as 3G, with a strategy for survival.

These arguments should not be misconstrued as favoring noninteroperability as a general matter. On the contrary, I would expect most network owners to continue to adhere to a basic architecture based TCP/IP. Maintaining interoperability provides consumers and network owners with such substantial financial advantages that most will adopt standardized protocols voluntarily. In most cases, then, mandating network neutrality would be superfluous. The only situations in which network neutrality has any purpose are those in which the market exhibits a preference for nonstandardization. My concern is that compelling interoperability under those circumstances runs the risk of reducing economic welfare, either by preventing the realization of

[85] The discussion that follows expands upon ideas I initially advanced in a brief editorial. *See* Christopher S. Yoo, *Fighting Traffic on the Disinformation Superhighway*, NASHVILLE TENNESSEAN, July 8, 2003, at 7.

efficiencies or by reinforcing the economies of scale that are the primary causes of potential market failure.

A. Economic Efficiencies from Vertical Integration

In addition to finding common ground on the structural preconditions necessary for vertical integration to harm competition, both Chicago and post-Chicago School theorists agree that vertical integration can yield substantial cost efficiencies.[86] The potential for enhanced economic welfare from vertical integration is reflected in the Merger Guidelines, which explicitly recognize that the efficiencies created from vertical merger may outweigh the possibility of anticompetitive effects.[87]

The broadband industry possesses many characteristics that make it likely that allowing a greater degree of vertical integration would yield substantial economic efficiencies.[88] For example, the presence of large, up-front fixed costs leave both network owners and content/application providers vulnerable to a range of opportunistic behavior that vertical integration can substantially mitigate. In addition, the fact that last-mile broadband providers must necessarily maintain a packet-switched network within their primary facilities to hold the data-based traffic after it has been separated from the other forms of communications[89] makes it unsurprising that last-mile broadband providers find it more economical to provide ISP services themselves.

The presence of such efficiencies is perhaps demonstrated most dramatically by the manner in which the multiple ISP access mandated during the AOL-Time Warner merger has been implemented.[90] Contrary to the original expectations of the FTC, the unaffiliated ISPs that have obtained access to AOL-Time Warner's cable modem systems under the FTC's merger clearance order have not placed their own packet network and backbone access facilities within AOL-Time Warner's headends. Instead, traffic bound for these unaffiliated ISPs

[86] *See* Yoo, *supra* note 27, at 192-200 (reviewing efficiencies resulting from vertical integration identified by Chicago School commentators); *id.* at 204 (reviewing the acknowledgement by post-Chicago theorists that vertical integration can yield substantial efficiencies).

[87] *Non-Horizontal Merger Guidelines*, *supra* note 76, §§ 4.135, 4.24. In addition, the Guidelines give more weight to expected efficiencies in the case of vertical integration than with respect to a horizontal merger. *Id.* § 4.24.

[88] For a more detailed analysis of this phenomenon, see Yoo, *supra* note 27, at 260-64. *See also* Joseph Farrell & Philip J. Weiser, *Modularity, Vertical Integration and Open Access Policies: Towards a Convergence of Antitrust and Regulation in the Internet Age*, 17 HARV. J.L. & TECH. 85, 97-105 (2004).

[89] *See supra* Section I.B.

[90] *See* Spulber & Yoo, *supra* note 16, at 1023 n.728.

exits the headend via AOL-Time Warner's backbone and is handed off to the unaffiliated ISP at some external location. It is hard to see how consumers benefit from such arrangements, given that they necessarily use the same equipment and thus provide the same speed, services, and access to content regardless of the identity of their nominal ISP.[91] The fact that these unaffiliated ISPs have found it more economical to share AOL Time Warner's existing ISP facilities rather than build their own strongly suggests that integrating ISP and last-mile operations does in fact yield real efficiencies.

The absence of consumer benefits underscores the extent to which compelled access represents something of a competition policy anomaly.[92] When confronted with an excessively concentrated market, competition policy's traditional response is to deconcentrate the problematic market, either by breaking up the existing monopoly or by facilitating entry by a competitor. Compelled access, in contrast, leaves the concentrated market intact and instead simply requires that the bottleneck resource be shared. Such an approach may be justified if competition in the concentrated market is infeasible, as was generally believed to be the case with respect to local telephone service until recently. Simply requiring that the monopoly be shared is inappropriate when competition from new entrants is technologically and economically achievable.[93]

B. The Tradeoff Between Network Standardization and Product Variety

The current debate has largely ignored how network neutrality can harm economic welfare by limiting the variety of products. The predominance of price theory, in which the sole source of economic welfare is the difference between reservation prices and the actual prices charged, has caused commentators studying the economics of broadband networks to overlook the potential benefits associated with product

[91] *See* COLUMBIA TELECOMMUNICATIONS CORPORATION, TECHNOLOGICAL ANALYSIS OF OPEN ACCESS AND CABLE TELEVISION SYSTEMS 22-23 (Dec. 2001), *available at* http:// archive.aclu.org/issues/cyber/broadband_report.pdf.

[92] *See* Yoo, *supra* note 27, at 268-69; Spulber & Yoo, *supra* note 16, at 1020.

[93] The feasibility of platform competition underscores the problems with viewing previous efforts to standardize and compel access to the local telephone service as precedent for imposing network neutrality on the Internet. *See* LESSIG, *supra* note 5, at 147-51; Lemley & Lessig, *supra* note 5, at 934-36, 938. Most steps to mandate access to local telephone networks were justified by the fact that competition in local telephony was believed impossible at the time. Such arguments do not apply to broadband, in which platform competition has emerged as a real possibility.

differentiation.[94] Simply put, allowing network owners to employ different protocols can foster innovation by allowing a wider range of network products to exist. Conversely, compulsory standardization can reduce consumer surplus by limiting the variety of products available.[95]

Viewed from this perspective, the pressure towards proprietary standards may not represent some sinister attempt by last-mile providers to harm competition. Instead, it may represent nothing more than the natural outgrowth of the underlying heterogeneity of consumer preferences. In the words of two leading commentators on network economics, "market equilibrium with multiple incompatible products reflects the social value of variety."[96] It is for this reason that economic theorists have uniformly rejected calls for blanket prohibitions of exclusivity arrangements and other means for differentiating network services.[97] Indeed, some models indicate that the deployment of proprietary network standards may actually prove more effective in promoting innovation and the adoption of socially optimal technologies.[98]

The current forces that are motivating network providers to consider introducing increasing levels of intelligence into their core networks provide an apt illustration of this dynamic. As discussed

[94] *See* Christopher S. Yoo, *Copyright and Product Differentiation*, 79 NYU L. REV. 212, 236-46 (2004) [hereinafter Yoo, *Copyright and Product Differentiation*] (reviewing the literature on product differentiation); Christopher S. Yoo, *Rethinking the Commitment to Free, Local Television*, 52 EMORY L.J. 1579, 1602-18 (2003) [hereinafter Yoo, *Rethinking Free, Local Television*] (applying product differentiation theory to electronic communications).

[95] *See, e.g.*, Michael L. Katz & Carl Shapiro, *Systems Competition and Network Effects*, 8 J. ECON. PERSP. 93, 110 (1994) (noting that "the primary cost of standardization is loss of variety: consumers have fewer differentiated products to pick from"); Farrell & Saloner, *supra* note 7, at 71 (counting "reduction in variety" as one of the "important social costs" of standardization).

[96] Katz & Shapiro, *supra* note 95, at 106 (citing Joseph Farrell & Garth Saloner, *Standardization and Variety*, 20 ECON. LETTERS 71 (1986)); *see also* S. J. Liebowitz & Stephen E. Margolis, *Should Technology Choice Be a Concern of Antitrust Policy?*, 9 HARV. J.L. & TECH. 283, 292 (1996) ("Where there are differences in preference regarding alternative standards, coexistence of standards is a likely outcome."); James B. Speta, *A Vision of Internet Openness by Government Fiat*, 96 NW. U. L. REV. 1553, 1569 (2002) ("If there were competition among broadband platforms, companies would pursue different strategies to differentiate themselves . . . .").

[97] *See, e.g.*, David Balto, *Networks and Exclusivity: Antirust Analysis to Promote Network Competition*, 7 GEO. MASON L. REV. 523 (1999); David S. Evans & Richard Schmalensee, *A Guide to the Antitrust Economics of Networks*, 10 ANTITRUST 36 (1996); Carl Shapiro, *Exclusivity in Network Industries*, 7 GEO. MASON L. REV. 673, 678 (1999).

[98] *See* Michael L. Katz & Carl Shapiro, *Product Introduction with Network Externalities*, 40 J. INDUS. ECON. 55, 73 (1992); Michael L. Katz & Carl Shapiro, *Technology Adoption in the Presence of Network Externalities*, 94 J. POL. ECON. 822, 825, 838-39 (1986).

earlier,[99] consumer demand for more time-sensitive applications, such as Internet telephony and streaming media, may be providing much of the impetus away from standardization. Forbidding network owners to introduce routers that can assign different priority levels to packets based on the nature of the application would have the effect of precluding consumers from enjoying the benefits of certain types of applications.[100] The current ubiquity of TCP/IP makes it seem like an appropriate default rule and appears to justify placing the burden on those who would deviate from it. A moment's reflection makes clear how adherence to the Internet's nonproprietary structure may actually impede innovation.

There is considerable irony in the network neutrality proponents' insistence that allowing Internet providers to introduce intelligence into their core networks would skew innovation and that technological humility demands adherence to an end-to-end architecture. The decision to concentrate intelligence at the edges of the network and to require packet nondiscrimination would itself skew the market towards certain applications and away from others. The choice is thus not between neutrality and nonneutrality in the overall direction of innovation. Mandating either would have the inevitable effect of determining technological winners and losers. My point is not that policy makers should reverse the presumption and erect a preference for innovation in the network's core over innovation at the network's edge. The better course is to favor neither and to allow consumer preferences to dictate the eventual outcome.

Some of the more thoughtful network neutrality proponents concede that consumers may well benefit from allowing broadband network owners to deploy proprietary protocols and that it can be difficult, if not impossible, to distinguish whether procompetitive or anticompetitive motivations prompted a particular network owner's conduct.[101] In light of the ambiguity regarding the economic impact of any particular use of proprietary protocols, it is somewhat surprising that network neutrality proponents nonetheless turn to government-mandated uniformity as their preferred regulatory response. The difficulties in distinguishing legitimate business practices from those motivated by a desire to harm competition would appear to favor the adoption of a contextual, case-by-case methodology over the use of categorical regulatory mandates.[102] Moreover, the position advanced by

[99] *See supra* Section I.B.

[100] *See* Speta, *supra* note 96, at 1574.

[101] *See* LESSIG, *supra* note 5, at 46-47, 167-76; Cooper, *supra* note 5, at 1050-52; Wu, *supra* note 5, at 148.

[102] For a related proposal, see Weiser, *supra* note 24, at 48-57 (advocating a case-by-case regulatory approach that erects a presumption against discriminatory access, but allows the

network neutrality proponents implicitly assumes that the government is in a better position to evaluate the competitive impact of particular practices than are private individuals and that the benefits of governmental intervention will outweigh the inevitable costs imposed by a regulatory lag.[103] That network neutrality advocates would embrace such a position is rendered all the more puzzling by the fact that it contradicts the decentralized, nonhierarchical spirit that they claim has animated the Internet since its inception.[104]

C. Network Neutrality and Competition in the Last Mile

On a more fundamental level, network neutrality advocates' focus on innovation in content and applications may be misplaced. Application of the basic insights of vertical integration theory reveals that markets will achieve economic efficiency only if each stage of production is competitive.[105] In other words, any vertical chain of production will only be as efficient as its most concentrated link. The central focus of broadband policy should be on how best to foster competition in the last mile. The intuition underlying this insight can be easily discerned by returning to the thought experiment in which we supposed that regulators required complete vertical disintegration of the broadband industry. As noted earlier, the fundamental economic problems stemming from the paucity of last-mile options would persist until new entrants appear.

Viewing the issues in this manner reveals how the major network neutrality proposals are focusing on the wrong policy problem. By directing their efforts towards encouraging and preserving competition among ISPs and content/application providers, they concentrate their attention on the segments of the industry that are already the most competitive and the least protected by entry barriers.[106] Restated in terms of the "layered model" of the broadband industry, the major network neutrality proposals advocate regulating the logical layer in a

network owner to offer legitimate business reasons to justify the practice).

[103] *See* Philip J. Weiser, *The Internet, Innovation, and Intellectual Property Policy*, 103 COLUM. L. REV. 534, 581 (2003).

[104] See Lessig, supra note 5, at 37, 40, 44. I must confess to being somewhat skeptical of the historical claim that the essence of the Internet has been its freedom from centralized control. The supposedly libertarian Internet of 1995 was largely the product of direct governmental support provided by DARPA and the National Science Foundation. Conversely, the supposedly sinister forces pushing the Internet away from its interoperable structure are actually the result of the shift to private ordering. It would thus be quite ironic to support governmental intervention as a means for promoting decentralization and the lack of hierarchy.

[105] Yoo, *supra* note 27, at 241-42.

[106] *See TCI-AT&T Merger*, *supra* note 28, at 3206 ¶ 93 (noting the high level of competition among ISPs).

way that best promotes competition in the application and content layers.[107] Broadband policy would be better served if such efforts were directed towards identifying and increasing the competitiveness of the most concentrated level of production. In other words, the logical layer should be regulated in the way that best promotes investment and the emergence of competition in the alternative physical network capacity, since it is the physical layer that is currently the most concentrated.

The lack of competition in the last mile has traditionally been attributed to both supply-side and demand-side considerations. The supply-side consideration is the fact that building the physical network of wires needed to provide DSL and cable modem service requires incurring substantial sunk costs. The presence of high sunk costs gives rise to a tendency towards natural monopoly conditions. The demand side consideration focuses on economic effects, which exist when the value of a network is determined by the number of other people connected to that network. The more people that are part of the network, the more valuable the network becomes. This dynamic can in turn create considerable demand-side economies of scale that reinforce the tendency towards concentration.

What has been largely overlooked in the current debates is how allowing networks to differentiate in the services they offer can mitigate the forces that are driving the broadband industry towards concentration. Conversely, measures that limit networks' ability to differentiate their services can exacerbate the already extant tendencies towards oligopoly in the last mile. There is thus a real possibility that imposing network neutrality may actually worsen, rather than alleviate, the central policy problem confronting the broadband industry.

1. Declining Average Costs and Supply-Side Economies of Scale

The supply-side considerations that cause last-mile services to exhibit a tendency towards natural monopoly can most easily be understood by focusing on the shape of the average cost curve.[108] If the

[107] *See, e.g.*, *Ex parte* Letter of Timothy Wu and Lawrence Lessig, *supra* note 40, at 2-9; Werbach, *supra* note 71, at 65-66.

[108] A more complete analysis of natural monopoly would require additional refinements. For example, a market may exhibit a tendency towards a natural monopoly even when average costs are increasing so long as the industry costs are subadditive, which occurs when one firm could produce the industry's entire output more cheaply than could two firms. WILLIAM J. BAUMOL ET AL., CONTESTABLE MARKETS AND THE THEORY OF INDUSTRY STRUCTURE 16-24 (rev. ed. 1988). That said, declining average costs are sufficient to give rise to natural monopoly. *Id.* at 176. *See generally* Yoo, *Rethinking Free, Local Television*, *supra* note 94, at 1596-1600 (discussing the determinants of declining average cost and their

average cost curve is decreasing, firms with the largest volumes can provide services the most cheaply, which in turn allows them to undercut their smaller competitors. This price advantage allows the largest players to capture increasingly large shares of the market, which reinforces their pricing advantage still further. Eventually the largest firm will gain a sufficient cost advantage to drive all of its competitors out of the market.

Whether average cost is increasing or decreasing is determined by the magnitude of the sunk costs. On the one hand, the ability to spread sunk costs over increasingly large volumes places downward pressure on average cost. For example, spreading a $100 million sunk-cost investment across one million customers would require allocating an average of $100 in sunk costs to each customer. If the same sunk-cost investment were spread over ten million customers, each consumer would have to pay only an average of $10 in order to cover sunk costs. The larger the sunk costs relative to the overall demand, the more pronounced these scale economies will be, although the marginal impact of this effect will decay exponentially as production increases. At the same time, the scarcity of factors of production and the principle of diminishing marginal returns tend to cause average costs to increase as volume increases.

Whether average cost is rising or falling at any particular point is determined by which of these two effects dominates the other. When the sunk-cost investments needed to establish the network are large, the former effect tends to loom as the more important and cause average cost to decline. Because entry by new broadband networks tends to require large sunk-cost investments, the market for last-mile providers is generally expected to exhibit a natural tendency towards concentration.

What network neutrality advocates have failed to recognize is how allowing last-mile broadband providers to differentiate their product offerings can help prevent declining-cost industries from devolving into natural monopolies.[109] It is not unusual for small-volume producers to survive against their larger rivals even in the face of unexhausted economies of scale by targeting those customers who place the highest

relationship to natural monopoly); Yoo, *Copyright and Product Differentiation*, *supra* note 94, at 226-28 (same).

[109] The seminal analysis of how competition among differentiated products can yield an equilibrium in which multiple declining-cost firms can coexist is EDWARD CHAMBERLIN, THE THEORY OF MONOPOLISTIC COMPETITION (7th ed. 1956). For more complete analysis of how product differentiation can mitigate the problems caused by declining average costs, see Yoo, *Copyright and Product Differentiation*, *supra* note 94, at 248-49. For a brief statement of how nonstandardization can facilitate competition among telecommunications networks, see Paul L. Joskow & Roger G. Noll, *The Bell Doctrine: Applications in Telecommunications, Electricity, and Other Network Industries*, 51 STAN. L. REV. 1249, 1251 (1999). For a discussion applying a similar analysis to another type of electronic communications, see Yoo, *Rethinking Free, Local Television*, *supra* note 94, at 1603 & n.61.

value on the particular types of products or services they offer, as demonstrated by the survival of specialty stores in a world increasingly dominated by larger and more efficient stores offering one-stop shopping. It is true that consumers of these small-volume producers will pay more for these specialized products. That said, it is difficult to see how these consumers are worse off. The value that they derive from the specialized product necessarily exceeds the amount they must pay for it, otherwise they simply would not agree to the transaction. Indeed, if consumers were unable to use higher prices to signal the intensity of their preferences, the low-volume version would not exist at all.

Last-mile providers have a number of avenues open to them for differentiating the networks. One way is by entering into exclusivity arrangements with respect to content, as demonstrated by the role played by such arrangements in helping direct broadcast satellite (DBS) provider DirecTV emerge as a viable alterative to cable television. For example, DirecTV is offering an exclusive programming package known as "NFL Sunday Ticket" that allows sports fans to watch the entire NFL schedule and not just the games being shown by CBS and Fox in their service area. Many cable customers have been frustrated by their inability to purchase NFL Sunday Ticket through their local cable operators. If regulators viewed exclusivity arrangement solely in static terms, they might be tempted to increase consumer choice by requiring this programming package also be made available to cable subscribers. The impolicy of such a reaction becomes manifest when one recalls that the central problem confronting the television industry is the local cable operators' historic dominance over multichannel video distribution. The market reaction has already demonstrated how exclusive programming like NFL Sunday Ticket is serving as a major driver towards the deployment of DBS as an alternative outlet for distributing television programming. Conversely, requiring that such programming be made available to cable as well as DBS customers would run the risk of further entrenching the local cable operator by eliminating one of the primary inducements to shift from cable to DBS.

Another way that last-mile providers can differentiate the services they provide is by optimizing the architecture of their networks for different types of applications. To offer an illustration in the context of broadband, it is theoretically possible that multiple broadband networks could co-exist notwithstanding the presence of unexhausted economies of scale. The first network could be optimized for conventional Internet applications, such as e-mail and website access. The second network could incorporate security features designed to appeal to users focused on e-commerce. The third network could employ policy-based routers that prioritize packets in the manner that allows for more effective provision of time-sensitive applications such as Internet telephony. Other networks could be designed to optimize the provision of still other services. If this were to occur, the network with the largest number of customers need not enjoy a decisive price advantage. Instead, each could survive by targeting and satisfying those consumers who place the highest value on the types of service they offer.

This example illustrates how imposing network neutrality could actually frustrate the emergence of platform competition in the last mile. Put another way, protocol standardization tends to commodify network services. By focusing competition solely on price, it tends to accentuate the pricing advantages created by declining average costs, which in turn reinforces the market's tendency towards concentration. Conversely, increasing the dimensions along which networks can compete by allowing them to deploy a broader range of architectures may make it easier for multiple last-mile providers to co-exist.[110]

2. Network Externalities and Demand-Side Economies of Scale

Other commentators have argued that network neutrality must be mandated as a regulatory matter in order to redress the competitive problems posed by network economic effects.[111] For reasons that I have discussed in detail elsewhere,[112] such claims are subject to a number of important analytical limitations and qualifications. A few brief comments on two of the more salient limitations will suffice to make my point.

First, for reasons analogous to the similar requirement with respect to vertical integration, the existing theories require that the network owner have a dominant market position before network economic effects can even plausibly harm competition.[113] The classic illustration of this

[110] By emphasizing the promotion of platform competition, my argument bears some resemblance to the proposal advanced by Philip Weiser. *See* Weiser, *supra* note 103, at 583-91. Our analyses differ in that Professor Weiser focuses his attention on the application and logical layers of the Internet, see *id.* at 542, whereas I am primarily concerned with competition in the physical layer. We also differ in our preferred policy response to a dominant player. Professor Weiser would support allowing others to have access to a proprietary protocol if the protocol owner achieves or is headed towards a dominant position. *Id.* at 591-94. I would attempt to dispel dominance not by direct regulation, but rather by attempting to facilitate entry by new broadband platforms and allowing the ensuing competition to dissipate any problems. Thus, my analysis favors allowing the use of proprietary protocols even when one firm is dominant. It also has the advantage of charging regulators with tasks for which they are better suited and establishing a regime that envisions an end to governmental intervention.

[111] *See supra* note 8 and accompanying text.

[112] *See* Yoo, *supra* note 27, at 278-82; Spulber & Yoo, *supra* note 16, at 924-33.

[113] Spulber & Yoo, *supra* note 16, at 923, 926.

phenomenon is the development of competition in local telephony during the 1890s made possible by the expiration of the initial telephone patents. After the Bell System's market share was cut in half, it attempted to employ network economic effects to reverse its losses. Specifically, it refused to interconnect with these upstarts, hoping that its greater network size would make it sufficiently more attractive to consumers to give it a decisive advantage. This effort ultimately failed, however, since the independent companies that comprised the other half of the industry were able to forestall any negative network economic effects by allying with one another to form a network that was similar in size to the Bell network.[114] In the end, it was control of certain patents critical to providing high-quality long distance service and not network economic effects that allowed the Bell System to return to dominance. The clear implication is that the presence of a single competitor of roughly the same size as the network owner is likely sufficient to eliminate any such problems.

Second, the argument that network economic effects create externalities that lead to market failure is wholly inapplicable in the context of telecommunications networks.[115] This is because any externalities that may exist will necessarily occur within a physical network that can be owned.[116] Thus, although individual users may not be in a position to capture all of the benefits created by their demand for network services, the network owner will almost certainly be in a position to do so. Any benefits created by network participation can thus be internalized and allocated through the interaction between the network owner and network users.[117]

The commentary on network economic effects thus does not support the contention that imposing network neutrality is necessary to protect competition. Quite the contrary, the literature indicates that compelling interoperability could affirmatively harm competition. This is because allowing last-mile providers to differentiate their networks can mitigate the problems resulting from any demand-side economies of scale created by network economic effects that may exist. Simply put, allowing networks to tailor their services to the needs of different groups

[114] *See* Roger Noll & Bruce M. Owen, *The Anticompetitive Uses of Regulation:* United States v. AT&T, *in* THE ANTITRUST REVOLUTION 290, 291-92 (John E. Kwoka, Jr. & Lawrence J. White eds., 1989).

[115] The discussion that follows is based on Spulber & Yoo, *supra* note 16, at 926-27.

[116] The literature refers to network externalities that occur in the context of a physical network as "direct network externalities." Katz & Shapiro, *supra* note 7, at 424.

[117] *See* S. J. Liebowitz & Stephen E. Margolis, *Are Network Externalities a New Source of Market Failure?*, 17 RES. LAW & ECON. 1, 11-13 (1995); S. J. Liebowitz & Stephen E. Margolis, *Network Externality: An Uncommon Tragedy*, 8 J. ECON. PERSP. 133, 137, 141-44 (1994).

of customers can offset the economic advantages enjoyed by larger networks in much the same manner as differentiation can offset the supply-side economies of scale. Targeting those customers who value the differentiated services makes it possible for smaller networks to survive despite the greater inherent appeal of larger networks.[118]

Conversely, mandating that all broadband networks employ nonproprietary protocols can foreclose network owners from using differentiation to mitigate the pressures towards concentration. Preventing network owners from varying the services that they offer forces networks to compete solely on price and network size, further reinforcing and accentuating the benefits already enjoyed by the largest players. As a result, network neutrality runs the danger of becoming the source of, rather than the solution to, market failure, thus allowing less innovation and fewer participants.

V. THE ROLE OF REGULATION

It is thus clear that permitting last-mile providers to deviate from the universal interoperability envisioned by the proponents of network neutrality may actually yield substantial economic benefits. Not only does differentiation potentially put networks in a better position to satisfy any underlying heterogeneity in consumer preferences; it also has the potential to alleviate the supply-side and demand-side economies of scale that are the sources of market failure that justify regulatory intervention in the first place.

The case against network neutrality is further bolstered by the risk that regulation might itself induce market failure by causing the existing oligopoly in last-mile technologies to persist long after technological improvements have made real competition possible. If access to a bottleneck network were not compelled, those who did not want to pay supracompetitive prices for network services would have the incentive to invest in alternative network capacity. Compelling access, on the other hand, would rescue those who would otherwise be financing the buildout of other last-mile technologies from having to undertake those investments. Network neutrality may thus have the effect of depriving alterative broadband platforms of their natural strategic partners and of starving them of the resources they need to build out their networks. Although such a policy might have been reasonable during previous eras, when the fact that construction of new network platforms was unfeasible rendered such considerations immaterial, it is unjustifiable in the current

[118] *See* Farrell & Saloner, *supra* note 96; Katz & Shapiro, *supra* note 95, at 106; Liebowitz & Margolis, *supra* note 96, at 292. For a related argument, see Weiser, *supra* note 103, at 587-89.

environment, in which competition from alternative network platforms is a real option.

The task confronting policy makers is made all the more difficult by the fact that making any difference would require policy makers to intervene at a fairly early stage in the technology's development, since governmental intervention after the market has settled on the optimal technology would serve little purpose.[119] Although whether regulation or private ordering would provide the better means for determining the optimal technology is ultimately an empirical question, there are a number of examples that suggest that public policy would be better served by relying on the latter. For example, during its early years the electric power industry went through an extended period of competition between standards based on direct current (DC) and alternating current (AC) that enhanced competition and promoted innovation in electrical appliances.[120] Even now, the electrical power network is diverse enough to accommodate appliances designed to run on the predominant 110-volt standard as well as larger appliances requiring 220 volts. Another example, drawn this time from the telecommunications industry, is the competition between time division multiple access (TDMA) and code division multiple access (CDMA) standards for mobile telephony. Rather than imposing a particular technological vision, the government has allowed these standards to compete in the marketplace.

In addition, governmental processes are subject to a number of well-recognized biases. Regulatory decisions are all too often shaped by political goals that are not always consistent with good policy.[121] In addition, policymakers may also find it tempting to give too little weight to the future benefits associated with the entry of alternative network capacity, which will no doubt seem uncertain and contingent, and to overvalue the more immediate and concrete benefits of providing consumers with more choices in the here and now. Indeed, the FCC has allowed short-term considerations to override longer-term benefits in the past.[122] Public choice theory strongly suggests that the bias in favor of

[119] Bresnahan, *supra* note 71, at 200-03.

[120] BRUCE M. OWEN & GREGORY L. ROSSTON, LOCAL BROADBAND ACCESS: *PRIMUM NON NOCERE* OR *PRIMUM PROCESSI*? A PROPERTY RIGHTS APPROACH 11-12 (AEI-Brookings Joint Center for Regulatory Studies Related Publication No. 03-19, Aug. 2003), *available at* http://www.aei.brookings.org/admin/authorpdfs/page.php?id=285 (citing Paul A. David & Julie Ann Bunn, *Gateway Technologies and the Evolutionary Dynamics of Network Industries: Lessons from Electricity Supply History*, *in* EVOLVING TECHNOLOGY AND MARKET STRUCTURE 121 (Arnold Heertje & Mark Perlman eds., 1990)). There is thus some irony in the fact that some network neutrality proponents point to the example of electric power as supporting the need for early governmental intervention. *See Ex parte* Letter of Timothy Wu and Lawrence Lessig, *supra* note 40, at 3; Wu, *supra* note 71, at 1165.

[121] *See* Bresnahan, *supra* note 71, at 202-03.

[122] *See* Christopher S. Yoo, *The Rise and Demise of the Technology-Specific Approach*

the former over the latter is no accident.[123]

There thus appears to be considerable danger that compelling access will forestall the buildout of 3G, fixed wireless, and other alternative broadband platforms.[124] I acknowledge the possibility that last-mile broadband providers may be able to use the market power provided by the degree of concentration in local markets to harm competition. For example, it is conceivable that cable operators might prohibit cable modem customers from streaming video in order to protect their market position in the market for conventional television. At the same time, such a prohibition might also represent an understandable attempt to prevent high-volume users from imposing congestion costs on other users.[125] Even network neutrality proponents acknowledge how difficult it can be to determine which is the case.[126]

In effect, policymakers are presented with a choice between two possible responses. On the one hand, they can trust their ability to distinguish between these two different situations and limit network neutrality to those in which deviations from full interoperability are motivated by anticompetitive considerations. The costs of doing so include the danger that regulators might err in making this determination as well as the risk that compelling access might delay entry by alternative last-mile technologies. On the other hand, regulators can adopt a more humble posture about their ability to distinguish anticompetitive from procompetitive behavior and attempt to resolve the problem by promoting entry by alternative broadband platforms. Once a sufficient number of alternative last-mile providers exist, the danger of anticompetitive effects disappears, as any attempt to use an exclusivity arrangement to harm competition will simply induce consumers to obtain their services from another last-mile provider. In this case, the

to the First Amendment, 91 GEO. L.J. 245, 272-75 (2003).

[123] There are also practical reasons to question the efficacy of access as a remedy. Network owners can be expected not to cooperate with those seeking access by charging the highest prices possible and by imposing restrictive nonprice terms and conditions. As a result, the FCC is likely to find itself embroiled in having to police all aspects of the parties' business relationship. This has led some scholars that suggest that attempts to mandate are likely to prove futile. *See* Paul L. Joskow & Roger G. Noll, *The Bell Doctrine: Applications in Telecommunications, Electricity, and Other Network Industries*, 51 STAN. L. REV. 1249 (1999). Indeed, the FCC's experience in implementing the UNE access requirements of the Telecommunications Act of 1996 appears to confirm this suspicion. *See also* Time Warner Entm't Co. v. FCC, 93 F.3d 957, 970 (D.C. Cir. 1996) (describing difficulties in implementing leased access to cable systems).

[124] *See* Yoo, *supra* note 27, at 268-69; Spulber & Yoo, *supra* note 16, at 1020; *see also* Glenn A. Woroch, *Open Access Rules and the Broadband Race*, 2002 L. REV. MICH. ST. U. DET. C.L. 719 (presenting a formal economic model of this effect).

[125] *See* James B. Speta, *The Vertical Dimension of Cable Access*, 71 U. COLO. L. REV. 975, 1004-07 (2000).

[126] *See supra* note 101 and accompanying text.

primary costs stem from delay. Because entry by new network platforms will not be instantaneous, there will necessarily be a period of time during which consumers may remain vulnerable to anticompetitive behavior.[127]

Choosing between these two approaches depends upon weighing their relative merits, with the understanding that each represents a second-best alternative. Although a formal analysis of the tradeoff exceeds the scope of my comments, my instinct is to favor the latter. It is motivated in part by my belief that regulatory authorities will be more effective at pursuing the goal of stimulating entry by new network platforms than they would be in ascertaining whether a particular exclusivity arrangement would promote or hinder competition. In addition, because the long-term benefits will be compounded over an indefinite period of time, they should dominate whatever short-run static inefficiency losses that may exist.[128] Perhaps most importantly, promoting entry has embedded within it a built-in exit strategy. Once a sufficient number of broadband network platforms exist, regulatory intervention will no longer be necessary. This stands in stark contrast with access-oriented solutions, which implicitly assume that regulation will continue indefinitely.

CONCLUSION

The claim that guaranteeing interoperability and nondiscrimination would benefit consumers has undisputed intuitive appeal. The fact that interoperability and neutrality have represented the historical norm makes it seem appropriate to put the burden of persuasion on those who would move away from that architecture.

A close examination of the economic tradeoffs underlying network neutrality reveals a number of countervailing considerations that may not be readily apparent at first blush. Not only does network neutrality risk reducing consumer choice in content and applications; it raises the even more significant danger of stifling competition in the last-mile by forestalling the emergence of new broadband technologies. Although such an admonition would be well taken under any circumstances, it carries particular force in industries like broadband that are undergoing rapid technological change.

[127] *See* Weiser, *supra* note 103, at 561; Yoo, *Copyright and Product Differentiation*, *supra* note 94, at 254 n.135.

[128] *See* Janusz Ordover & William Baumol, *Antitrust Policy and High-Technology Industries*, 4 OXFORD REV. ECON. POL'Y 13, 32 (1988); David J. Brennan, *Fair Price and Public Goods: A Theory of Value Applied to Retransmission*, 22 INT'L REV. L. & ECON. 347, 355 (2002).

THE BROADBAND DEBATE, A USER'S GUIDE

TIM WU[*]

INTRODUCTION

"What ever happened to 'Hands off the Net?' "[1]

Back in the 1990s, Internet communications policy was easier. It was easy to agree that the network's growth ought not be impended by excessive government regulation. It was easy to hope that the Internet would solve all of its own problems. Yet it turned out that the success of the network was hiding strong differences of opinion. Today, the euphoria is gone, and the divide in Internet communications policy has become clear and unmistakable. It most clearly a divide between two distinct groups: the self-proclaimed "Openists" and "Deregulationists."

This divide will do much to inform the reform of the Telecommunications Act in general, and Broadband policy in particular.

[*] Visiting Professor, Columbia Law School, Associate Professor, University of Virginia School of Law. My thanks to participants at the 2004 Silicon Flatirons conference, and the March 26, 2004 Policy Forum "The Future Of The Internet In The Broadband Age" organized by the Consumer Federation of America and the Stanford Center for Internet and Society. Specific thanks for comments from Phil Weiser and Mark Cooper, discussions with Jim Speta, Christopher Yoo and Mark Lemley, and to Lee Kovarksy for research assistance.

[1] Adam Thierer, *Congressional Tech Agenda for Rest of Year = Just More Regulation*, THE TECHNOLOGY LIBERATION FRONT (Sept. 7, 2004), *at* http://www.techliberation.com/archives/014257.php.

Accordingly, this Article is meant as a user's guide to the broadband policy debate: a guide to what separates us, and what might make reconciliation possible. It is optimistic that policy reconciliation is possible, though aware that saying so doesn't make it likely.

The summary of the debate is critical. I fault the Openists for being too prone to favor regulation without making clear the connection between ends and means. For example, too few Openists have asked the degree to which the structural "open access" remedies pushed by independent service providers actually promote the Openists' vision.[2] Meanwhile, I fault the Deregulationists for two reasons. First, the Deregulationists have overlooked the fact that limiting government, as they desire, sometimes *requires* government action. Remedies like network neutrality, for reasons I suggest, may be as important for control of government as of industry. I also fault the Deregulationists for an exaggerated faith in industry decision-making. I suggest that some Deregulationists have failed to familiarize themselves with the processes of industry decision-making before demanding deference to it. This is a particularly serious problem given that the telecommunications industry has a recent track record of terrible judgment and even outright fraud. An important example is the demand of some Deregulationists that deference is due to a so-called "smart pipe" vision, without analysis of whether that vision has any independent merit.

The article, finally, explores a reconciliation of the broadband debate with the network neutrality principle as a starting point. Deregulations and Openists, while divided along many lines, share a common faith in innovation as the basis of economic growth. Both sides, in short, worship Joseph Schumpeter and his ideas of competitive, capitalistic innovation. Fidelity to this shared faith should mean mutual surrender of idealized models of either government or powerful private entities, respectively, in exchange for a shared cynicism. We should recognize that both government and the private sector have an unhappy record of blocking the new in favor of the old, and that such tendencies are likely to continue.

Reconciliation, I (optimistically) believe, is possible. The Deregulationist and Openist ought remember their common dedication to a single principle: free and unmediated market entry, symbolized by the rubber-cup of Hush-A-Phone.[3] It is by returning to such points of consensus that the reconciliation of communications policy can begin.

[2] *See also* Tim Wu, *Network Neutrality, Broadband Discrimination*, 2 J. ON TELECOMM. & HIGH TECH. L. 141 (2003) (expanding on this point).

[3] *See* Hush-A-Phone Corp. v. United States, 238 F.2d 266, 269 (D.C. Cir. 1956) (holding that the FCC cannot block the attachment of reasonable network attachments,

I argue that neither Deregulationists or Openists should have reason to oppose Network Neutrality rules that create rights in users to use the applications or equipment of their choice. This is a position that many Deregulationists, including FCC Chairman Michael Powell, have come to endorse. What both sides should want in an inevitable regulatory framework for broadband are rules that pre-commit both industry and government to open market entry. It must be remembered that rules creating rights in users also guarantee the right of operators to enter the application market, free of government hindrance. For these and other reasons discussed below, limited network neutrality rules should on reflection be attractive to both sides.

Section I describes the emergent divide in the visions of the future that underlie today's policy divisions. Section II explains some of what unites and divides in the economics of the Deregulationists. Section III argues for broadband reconciliation premised on user rights to access the content, applications and equipment of their choice.

I. VISIONS OF THE FUTURE

Communications theorists, like everyone else, have their visions of an ideal future that drive more of their arguments than they would like to admit. While the theorist's utopia has much less sand and sunshine than the average person's, its importance is nonetheless axiomatic.

A. The Openists

In the communications world some technologies attract what you might call a high chatter to deployment ratio. That means the volume of talk about the technology exceeds, by an absurd ratio, the actual number of deployments. "Videophones" are a great historical example, as is "Video-on-Demand" and, of course, the glacial sixth version of the Internet protocol (IPv6). In the 1990s, the technology named Voice over IP (VoIP) was a starring member of this suspect class. The technology promises carriage of voice signals using Internet technology, an attractive idea, and in the 1990s and the early 2000s it was discussed endlessly despite minimal deployment.

The discussion usually centered on the question: when would broadband carriers deploy VoIP? And the answer was always, "not quite yet." There were reasons. Many within the industry argued that VoIP was not a viable technology without substantial network improvements. Engineers said that the Internet Protocol was too inconsistent to guarantee voice service of a quality that any customer would buy.

namely the "Hush-A-Phone" device that attached to a handset and insulated telephone conversations against background noise).

Industry regulatory strategists, meanwhile, were concerned that offering voice service would attract federal regulation like honey attracts bees. As for the Bell companies, the main Digital Subscriber Line (DSL) providers, there was always the problem of providing a service that might cannibalize the industry's most profitable service.

But everyone was watching the wrong companies, for where broadband operators were timid, a company named Vonage was brave. In late 2003 Vonage leapfrogged the broadband operators and began selling VoIP directly to large volumes of customers. Vonage did so not by cooperating with broadband operators but avoiding them. It sold a plug-in device: an actual telephone that connects directly into the network and provided phone service for a fraction of the normal cost. It is true that the quality of the Vonage connection was not, to a telecommunications engineer, strictly of the same quality as that available on a traditional phone network. Yet Vonage's quality was fine to an American people schooled by cell phones; its many users claim they cannot tell the difference. Vonage, offered what everyone said no one would buy, and became the Internet's success story of 2004.[4]

The Vonage story captures much of the Openist's vision of what the Internet revolution has meant for communications policy. Without Vonage, VoIP would have arrived on the carrier's schedule: later or perhaps never. Vonage shows why Openists see the nation's communications network important, first and foremost, as an *innovation commons*—a resource for innovators from anywhere to draw upon.[5] The Openist credo is to care about the nation's communications infrastructure, not so much of itself, but for how it catalyzes the nation's economic and creative potential. Vonage was free to enter the market with a new way of selling voice service only because the network is open, its standards as "free as the air to common use."

The Openist's theory of an innovation commons can be broken into three prescriptive principles. The first is the *Infrastructure* principle. It is an insistence that the most important purpose of a communications network is as public infrastructure, with particular meaning attached to that concept. It means that the principal value of the network is indirect: it as a source of positive spillovers, or externalities, that enable the work of others. It suggests that the highest potential of the network will be achieved not by the accomplishments of network *owners* but by what creative users and developers can do with a fast and reliable connection

[4] *See* Stephen Wildstrom, *At Last, You Can Ditch The Phone Company VOIP Lets You Make Clear, Fast Calls Over The Net, Using A Plain Phone*, BUS. WK., May 17, 2004, at 26.

[5] *See, e.g.*, Lawrence Lessig, *The Internet Under Siege*, FOREIGN POL'Y, Nov. 1, 2001, *available at* http://www.lessig.org/content/columns/foreignpolicy1.pdf.

between every human on earth.

One way of understanding this vision of the network as "infrastructure" is to contrast it directly with its foil, the idea that a network is a "service" or "product" sold by a company. At the podium at the 2004 Silicon Flatirons Conference, speaker Mark Cooper put this *product / infrastructure* distinction in vivid terms:

> The proprietary platform folks are talking about a BETA Max, an Atari and an Xbox;
> I am talking a general purpose technology, a cumulative, systemic technology, like the railroad, electricity or the telephone.
>
> For them the end-to-end principle is an obscure garden variety interface;
> For me it is a fundamental design principle of an enabling technology.
>
> When they analyze the proprietary standards wars, there are few if any externalities;
> When I analyze a bearer service like the digital communications platform, externalities dominate.[6]

The second principle is the *Neutrality* principle. It holds that to reach its highest potential, a communications infrastructure must not discriminate as between uses, users, or content. As FCC Commissioner Michael Copps puts it: "From its inception, the Internet was designed, as those present during the course of its creation will tell you, to prevent government or a corporation or anyone else from controlling it. It was designed to defeat discrimination against users, ideas and technologies."[7]

The third principle is the *End-to-End* (e2e) principle. Whatever its meaning elsewhere,[8] in broadband policy e2e stands for a theory of

[6] Mark Cooper, Remarks at the Silicon Flatirons Telecommunications Program Conference, University of Colorado School of Law (Feb. 8, 2004) (transcript available from the Silicon Flatirons Telecommunications Program, http://www.silicon-flatirons.org) [*hereinafter* Cooper Remarks].

[7] *See* FCC Commissioner Michael J. Copps, The Beginning of the End of the Internet? Discrimination, Closed Networks, and the Future of Cyberspace, Address Before the New American Foundation (Oct. 9, 2003) (transcript available at http://www.newamerica.net/Download_Docs/pdfs/Docs_File_194_1.pdf).

[8] In the telecommunications industry, the term "end-to-end" is used for a variety of purposes, many of which are quite meaningless, or roughly synonymous with "good." *See, e.g.*, MOTOROLA, INC., MOTOROLA NEXT LEVEL COMMUNICATIONS, END-TO-END, *at* http://broadband.motorola.com/nlc/solutions/endtoend.asp (last visited Jun. 26, 2004). Christopher Yoo, meanwhile, writes in this volume that the end-to-end principle as originally described by the network engineering literature has been misunderstood by Openists. *See* Christopher S. Yoo, *Would Mandating Broadband Network Neutrality Help or Hurt Competition? A Comment on the End-to-End Debate*, 3 J. ON TELECOMM. & HIGH TECH. L. 23, 42-46 (2004).

innovation. It rejects centralized, planned innovation, and holds that the greatest rate of technological development is driven by delegating decisional authority to the decentralized "ends" of any network. The reason is fairly simple: the "ends" of the network are numerous, or nearly unlimited, and delegating authority to the ends opens the door to more approaches to a given technological challenge. The e2e principle assumes that innovation is an evolutionary process, driven by contests between competing approaches to a problem. For Openists, the e2e principle puts as many players in the contest as possible to ensure the true champion emerges.

Openists believe these three principles are what made the Internet different from other communications networks; they hold that the embedding of these principles in the design of the Internet is the essence of the revolution. Their founder's story rejects technological determinism, or the idea that the Internet was destined to occur. They instead see the founding engineers, men like Paul Baran, Vint Cerf and Robert Kahn, as heroic figures and communications revolutionaries.[9]

The Openist vision just described can seem abstract to regulators and policy-makers. For that reason, in recent years Openists have advanced a more concrete regulatory model to explain what neutrality would entail. That model suggests that the Internet will continue its success if we come to understand it as a more humble but nonetheless highly successful innovation enhancing network: the nation's electric grid.

While today taken for granted, the electric network is probably the greatest innovation catalyst of our age. The radio, the air conditioner, the computer and other giant innovations have all depended on a predictable and reliable supply of electric current.[10] This multipurpose network is like the railways of the 19th century or the first roads of ages past: among the foundations of the national economy.

Openists point to the electrical grid and say it is successful precisely because we don't care about electricity as a product, but care instead about what the electric grid makes possible. It provides a standardized platform for the development of appliances that serve human needs, such as the hair dryer or DVD player. Sony and IBM do business safe in the assumption that American electricity will be predictable, standardized, and provided without preference for certain brands or products. There is

[9] An example of the heroic version of the Internet's invention is KATIE HAFNER & MATTHEW LYON, WHERE WIZARDS STAY UP LATE, THE ORIGINS OF THE INTERNET (1996).

[10] The electric grid model appears in Mark Cooper's remarks at the Silicon Flatirons Conference. *See Cooper Remarks, supra* note 6; Tim Wu, *Application-Centered Internet Analysis*, 85 VA. L. REV. 1163, 1165 (1999); *see also* LAWRENCE LESSIG, CODE AND OTHER LAWS OF CYBERSPACE (2000).

no built-in favoritism for the VCR over the DVD player. You do not ask the electric companies permission before plugging in a new cordless phone. This makes the electric grid, Openists say, one of the greatest models of network neutrality the world has ever known.

The electric grid model returns us to the Vonage story that opened the section. The long term vision is a future where still other services long-centralized will finally be decentralized. Freestanding IP-televisions, IP-stereos, and many other services should be available based on plug-in devices, developed by independent, competing companies. This vision, in the Openists' view, is far from inevitable. It requires defense of the network against forces that want, for a variety of reasons, to close the network to market entrants.

B. The Deregulationists

The contrasting vision of the communications future begins with the decades-old idea of *media convergence*. Convergence means a natural technological progression toward a single network for communications services. Voice, data, and video, historically carried over different networks will, in the future, be carried over a single "pipe." There was a time, namely the 1990s, when twin visions of "convergence" and "commons" could maintain a peaceful coexistence. But today the visions are rivals, for the underlying principles are in conflict.

The convergence vision focuses on the *owners* of the networks and the services they will offer on the converged network "telecosm."[11] As Peter Huber puts it:

> Convergence among technologies is doing more than networking the networks. It is transforming the services; the vast capacities of broadband networks make nonsense of the traditional regulation distinction between "carriers" and "broadcasters." . . . Broadcasters, in short, are mastering the art of keeping the "broad" while switching the "cast." Telephone companies are keeping their switched, addressable capabilities while widening their bandwidth and their reach. Nobody casts drift nets anymore. They are all fly fishermen now.[12]

The Deregulationist position can also be reduced to several principles. First is the *Propertization* principle: any given resource will generally reach its best use when mapped out as property, and assigned owners. When Deregulationists think "commons," the word "tragedy" is never far

[11] The idea of a "telecosm" was described most vividly in GEORGE GILDER, TELECOSM (2000).

[12] PETER HUBER ET AL., FEDERAL BROADBAND LAW § 1.2.4 (1995).

from mind. Property owners can be expected to maintain and steward only what they have the right to exclude others from.[13] Additionally, the creation of transferable property rights will facilitate private, welfare-enhancing transactions. As Frank Easterbrook famously put it in *Cyberspace and the Law of the Horse*: "we need to bring the Internet into the world of property law . . . without which welfare-increasing bargains cannot occur."[14]

The second principle is the *Incentive* principle, which is just a simple reminder that communications networks are expensive investments and that companies will only build when given the prospect of a reasonable return on investment.[15] To speak, as Openists do, of a pure public infrastructure may have made some sense when the government was funding and building the network, but by now is seriously out-of-date. Some Deregulationists will accept that aspects of the Internet that have the character of a public good or natural monopoly and therefore might be best provided by an entity outside of the market (Internet addresses might be an example). But in general, and for most of the network and its applications, the private sector responding to appropriate incentives will drive and fund the future.

The final principle is *Deregulation* itself. The Deregulationist is naturally suspicious of government regulation outside of the assignment of property rights. This can be understood as a different interpretation of the Internet revolution: the greatest factor in the success of the Internet was the fact that the Commission and Congress largely stayed out of the way. The idea of technological destinies, discussed above, is important to this position. Deregulationists are generally technological realists, believing that power more than ideas determines the course of history. Government may slow but it cannot stop the inevitable. So while Openists may try to slow or stop it, in the long term the power of private network owners will drive the next-generation Internet.

Much of this is as abstract as the idea of an Internet commons. When asked for a more concrete vision of what Deregulationist policies may lead to, Deregulationists have turned to the vision of the "smart pipe." The smart pipe (also known as the "Quality of Service (QoS) Internet" or the "value-added service" model) is the central dogma of innumerable industry white papers. The basic idea is this: broadband operators will increase revenue and profit by selling applications bundled

[13] *Cf.* Garrett Hardin, *The Tragedy of the Commons*, 162 SCIENCE 1243 (1968).

[14] Frank H. Easterbrook, *Cyberspace and the Law of the Horse*, 1996 U. CHI. LEGAL F. 207, 212-13 (1996).

[15] *See, e.g.*, ADAM D. THIERER, "NET NEUTRALITY" DIGITAL DISCRIMINATION OR REGULATORY GAMESMANSHIP IN CYBERSPACE? (The CATO Institute, CATO Policy Analysis No. 507, Jan. 12, 2004), *available at* http://www.cato.org/pubs/pa507.pdf.

with a basic connection. Stated in industry jargon, broadband operators using "next-generation" technologies can offer their customers a host of "value-added" services, such as telephony, video-on-demand, and so on.[16] The incentive for this new model, at least on the authority of projection, is profits that far exceed what can be earned from selling "commodity bandwidth."

Equipment vendors have pushed this vision aggressively for the last decade. As a current Cisco White Paper instructs cable operators:

> Tomorrow's cable business growth, however, will come from offering value-added services to consumers such as video on demand (VOD), interactive TV, and cable telephony.[17]

How? As Cisco explains to cable operators, in a FAQ rich with industry jargon:

> The Cisco MSOC solution defines a multiservice network infrastructure for delivering HFC-based, revenue generating enhanced IP-based services. Cisco MSOC provides best-practice design guidelines for building a well-engineered, reliable, highly available and quality-of-service (QoS)-enabled cable network capable of supporting real-time sensitive applications (such as VoIP and commercial services). . . . The largely untapped market for enhanced IP-based services, beyond high-speed Internet, will primarily fuel the future revenue growth for the cable operators.[18]

In short, the vendor industry and Deregulationists predict that the next great wave of innovation will occur at the center of the network, not the ends.[19] That directly contradicts the end-to-end principle, but that's fine: most Deregulationists believe blind adherence to the end-to-end principle is what is in fact slowing technological progress today. Economists Bruce Owen and Gregory Rosston, for example, argue that "openness inevitably has a price," and that certain innovations "have been

[16] *See, e.g.*, Ira Brodsky, *Telecom Carriers Need to Smarter Up Their Pipes,* NETWORK WORLD FUSION, (Jan. 15, 2001), *at* http://www.nwfusion.com/columnists/2001/00280817.html.

[17] CISCO SYSTEMS, RESIDENTIAL CABLE SERVICES (2003), *available at* http://www.cisco.com/application/pdf/en/us/guest/netsol/ns289/c714/ccmigration_09186a008014e05f.pdf.

[18] CISCO SYSTEMS, MULTISYSTEM OVER CABLE SOLUTIONS (2003), *available at* http://www.cisco.com/en/US/netsol/ns341/ns396/ns289/ns269/netqa09186a0080113708.html.

[19] *See, e.g.*, BRUCE OWEN & GREGORY ROSSTON, LOCAL BROADBAND ACCESS: PRIMUM NON NOCERE OR PRIMUM PROCESSI? A PROPERTY RIGHTS APPROACH 21 (STAN. L. & ECON., Olin Working Paper No. 263, 2003), *available at* http://ssrn.com/abstract=431620.

slowed or even blocked because of the [e2e's] requirement that the network not have embedded intelligence."[20]

Finally, while Openists favor the story of the Internet founders, Deregulationists invoke a different prescriptive saga: the birth of cable television. As Peter Huber puts it "Cable was the prototype of the broadband future."[21] The development of the cable networks was a story of private ingenuity's victory over governmental perfidy and, in the mind of many Deregulationists, a story with clear lessons for broadband 2000.

The Commission in the 1960s was anxious to preserve certain ideal visions of television. The two most important were that it be free and that it be local. Whatever the theoretical merits of those views, Deregulationists point out that the practical effect was to slow the spread of cable television for a full decade and to stop it from penetrating urban markets.[22] It was only by the 1970s that the Commission finally relaxed its grip and let competitive forces run their course. (Today cable companies are the TV's dominant players, so much so that cable operators rather casually bid to acquire broadcasters, their one-time overlords.)[23]

This, the Deregulationists would suggest, is what's happening in broadband policy, though our proximity makes us incapable of realizing it. There are certain parallels that anchor the obstructionist story. First, physical broadband networks, whether cable, twisted pair, or wireless spectrum, are indeed the subject of intensely complex federal and state regulation, rather like those to which the cable industry was subjected in the late 1960s and early 1970s (one writer described the cable regulations of 1972 as the "most complicated scheme ever devised by the mind of man"[24]). The ongoing regulatory asymmetry of DSL, cable, and wireless services is perhaps the most obvious example of a governmentally introduced distortion.

Second, the Commission in this view is still attached to some inappropriately utopian visions, which do not correspond with technological destiny. Today, the Deregulationist would contend, replacing "localism" and "free television" are similarly impractical ideals

[20] *Id.* at 21.

[21] PETER HUBER, LAW AND DISORDER IN CYBERSPACE 62 (1997).

[22] *See* Leonard Chazen & Leonard Ross, *Federal Regulation of Cable Television, the Visible Hand*, 83 HARV. L. REV. 1820, 1820 (1970); Stanley M. Besen & Robert W. Crandall, *The Deregulation of Cable Television*, 4 LAW & CONTEMP. PROBS. 77, 94 (1981) ("Cable entered the 1970s as a small business relegated primarily to rural areas and small communities and held hostage by television broadcasters to the Commission's hope for the development of UHF.").

[23] See Alison Beard, *Comcast Must Spell Out Plan for ABC*, FIN. TIMES, Feb. 17, 2004, *available at* 2004 WL 70205529 (discussing Comcast's planned acquisition of ABC).

[24] Besen & Crandall, *supra* note 22, at 81-91 (documenting FCC activity constraining the growth of cable).

like the "end-to-end principle," "open access" and, of course "network neutrality."

A related similarity is what Deregulationists decry as an effort to prop up doomed businesses in the name of lofty ideals. In the 1960s, the Commission placed much hope for the future of television in a new generation of UHF broadcast stations.[25] UHF stations did have many appealing qualities: they were locally owned, free over the air for recipients, and available in greater quantity than VHF stations. But UHF was hopeless as a technological competitor to cable. Today, Deregulationists contend, we see the scenario repeating itself. Independent Internet Service Providers (ISPs) are kept alive in the vain hope that they may somehow make the broadband world a better place.

So what is the Deregulationist's vision of the future? Some argue that the FCC and Internet old-timers are holding back, not promoting the natural progress of broadband networks. Innovation, they contend, can happen anywhere, not just at the "ends." Dreams of a neutral network may be holding back the next communications revolution, one that will arise from the center of the network. That vision will necessarily be driven by private network owners and will bring consumers both what they want and are willing to pay for and what the old Internet could never have provided.

*

It is between substantive visions of the future where the Openist – Deregulationist divide is most stark. That is perhaps because the contrasting utopias depend mainly on intuition and aesthetics, and faith in the private and public sectors, respectively. Yet nonetheless the sides are not precise opponents. Openists are primarily focused on the ends—the innovation commons. Deregulations care most about the means, most of all wanting to prevent disastrous and long-lasting governmental intervention. There is room, in other words, for reconciliation.

[25] This was one of the arguments of the 1958 Cox Report. Kenneth Cox, The Problem of Television Service for Smaller Communities. *Staff Report to the Senate Committee on Interstate and Foreign Commerce*, 26 December 1958.

II. SHARED ECONOMIC FAITHS

A. Schumpeter

It is worth reemphasizing that the greatest unifying belief as between the Openist and Deregulationist is a common idolization of innovation. Both sides, with a few exceptions,[26] worship at the shrine of economist Joseph Schumpeter and admire his concept of innovation as "creative destruction."[27]

The core of what is agreed upon can be stated simply. Both sides take innovation, and not price competition, as the principle driver of economic growth. Proximity to the industries of high technology leads naturally to favoring or at least acknowledging what economists call "dynamic" economic models. Both the Openists and Deregulatists do not believe that reaching market equilibrium is a particularly attractive ideal: instead, new companies, new services and new products are the primary source of increased efficiency and economic growth. That belief, for both sides, put innovation policy at the center of national economic policy.

How, then, does innovation happen? As Schumpeter said, "Creative Destruction is the essential fact about capitalism."[28] Schumpeter's "capitalist" or "competitive" theory of innovation is centered on the "process of industrial mutation . . . that incessantly revolutionizes the economic structure from within, incessantly destroying the old one, incessantly creating a new one." Both sides also agree with Schumpeter that the greatest barrier to innovation is "ordinary routine." As he put it "knowledge and habit once acquired becomes as firmly rooted in ourselves as a railway embankment in the earth."[29] As a result, even "in the breast of one who wishes to do something new, the forces of habit raise up and bear witness against the embryonic project." The

[26] There is a dissenting Openist viewpoint that sees the value of open infrastructure primarily in terms of providing positive social externalities as opposed to for its role in spurring innovation. (We value open parks for walking and socializing, not because they lead to new inventions—the same should go for the Internet). This view is well expressed in Brett M. Frischmann, An Economic Theory of Infrastructure and Sustainable Infrastructure Commons (2004) (working manuscript, on file with author).

[27] Much as Schumpeter admired Karl Marx. See JOSEPH A. SCHUMPETER, CAPITALISM, SOCIALISM, AND DEMOCRACY 61 (1950) [*hereinafter* SCHUMPETER, CAPITALISM, SOCIALISM, AND DEMOCRACY] ("Can capitalism survive? No. I do not think it can"). Most of his account of capitalism as a system of growth through innovation as opposed to price competition is summarized in Ch. VII. *Id.*

[28] *Id.* at 83.

[29] JOSEPH A. SCHUMPETER, A THEORY OF ECONOMIC DEVELOPMENT 84 (1961) [*hereinafter* SCHUMPETER, ECONOMIC DEVELOPMENT].

greatest threat is social resistance, particularly from "the groups threatened by the innovation."[30]

As I said, most Openists and Deregulationists consider themselves Schumpeterians. With all this agreement, where do the differences arise? The difference between Openists and Deregulationists in Schumpeterian terms is over who the agents of creative destruction are. It boils down to something quite simple: the two sides have different attitudes toward size. Many Deregulationists, like the later Schumpeter, see large and powerful companies as the central agents of creative destruction. Big firms are the winners, the success stories, the smartest and strongest. For the Openists, conversely, size is not necessarily a sign of continuing success but instead suggestive of some knack for blocking market entry. The Openists like the early Schumpeter, and his younger focus on the entrepreneur as the seed of creative destructive. The difference in opinion over size can be as intractable as how one sees Sport Utility Vehicles or modern skyscrapers. Some see a mighty work of man, others see a wasteful monster. Yet since Schumpeter himself managed to reconcile the role of large and small in his work, it ought be possible for his latter-day followers.

First, the vision of the Deregulationists' Schumpeter: "What we have got to accept" said Schumpeter in 1943, is that the "large-scale establishment" is "the most powerful engine of [economic] progress and in particular of the long-run expansion of total output."[31] Putting faith in "perfect competition" among numerous competitors was, in his view, folly, for "the firm of the type that is compatible with perfect competition is in many cases inferior in internal, especially technological, efficiency."[32]

The reasons for this belief can be specified more carefully. First, in a dynamic market, when a firm successfully establishes a new market through product innovation, the result is inevitably at least a short-term market advantage, even a monopoly. Yet that market power is no cause for concern, as it will erode quickly under the pressure of capitalistic competition. Indeed, short-term monopoly profits are not a social ill but rather social boon. For it is the very existence of potential monopoly profit that fires the pistons of creative destruction. It is only the possibility of a giant and seemingly unfair payoff that motivates risky and otherwise irrational innovative behavior. Under Capitalism, Schumpeter said, "spectacular prizes much greater than would have been necessary to call forth the particular effort are thrown to a small minority of winners,

[30] *Id.*

[31] SCHUMPETER, CAPITALISM, SOCIALISM & DEMOCRACY, *supra* note 27, at 106.

[32] *Id.*

thus propelling much more efficaciously than a more equal and more 'just' distribution would."[33]

Second, large, powerful firms have advantages that in this view make them the only entities truly capable of producing meaningful progress. One idea, not strictly Schumpeterian, is that the large firm with a secure market may carry out product innovation in a planned and careful way, and decrease the waste from competing innovative processes.[34] Another idea from Schumpeter is that large firms are simply smarter, stronger, and better. Schumpeter argued that "there are superior methods available to the monopolist," and that "monopolization may increase the sphere of influence of the better, and decrease the sphere of influence of inferior brains."[35]

In the broadband context, this vision sees the great firms—mainly, the greatest of cable operators and powerful Bell Operating Companies—as the agents of perpetual revolution. Their battle for the giant profits that await the champion, the single broadband monopolist, are the driving force behind broadband innovation and the future of the Internet.

The Openists reject or temper this "naive" faith in great firms, both with the work of Schumpeter himself, and that of later evolutionary economists. Consider first the early, German-language Schumpeter who spent his time on individual entrepreneurs, and the challenges they face.[36]

Openists think that many have misunderstood Schumpeter: that he didn't truly believe that the large firm had an inherent advantage over the small firm. As economist Jan Farberberg argues, "In fact, Schumpeter seemed to be much more concerned with the difference between new and old firms than between small and large firms."[37] Meanwhile, the early Schumpeter's theory of entrepreneurs is distinct and compelling. They are to him unusual characters, risk-seeking individuals with a "special quality," who are spread through the population like undercover superheroes. What distinguished this class of individuals, said Schumpeter (foreshadowing the "open source" movement), was that profit would be but one motive and not the most important one. Instead, the entrepreneur was generally driven by "the dream or will to

[33] *Id.*

[34] *Cf.* Edmund W. Kitch, *The Nature and Function of the Patent System*, 20 J.L. & ECON. 265 (1977).

[35] SCHUMPETER, CAPITALISM, SOCIALISM & DEMOCRACY, *supra* note 27, at 101

[36] *See* SCHUMPETER, ECONOMIC DEVELOPMENT, *supra* note 29.

[37] JAN FAGERBERG, A LAYMAN'S GUIDE TO EVOLUTIONARY ECONOMICS 15 (Centre for Technology, Innovation and Culture, Oslo, TIK Working Paper, Sept. 2002), *available at* http://folk.uio.no/janf/downloadp/02fagerberg_evolution.pdf.

found a private kingdom;" "the will to conquer: the impulse to fight, to prove oneself superior to others" and finally the "joy of creating."[38]

The Openist also directs those of Schumpeterian faith to the work of recent evolutionary economists like Richard Nelson and Sidney Winter. An essential element of such neo-Schumpeterian work is the emphasis on the uncertainty and contingency and of technological outcomes. It predicts multiple possible equilibria, rather than a single, predictable outcome. One reason is that this branch of economic thinking takes a much more sophisticated view of how firms decide what to do, rejecting the premise that firms will generally arrive at "maximizing" decisions.[39] Firms instead generally depend on a set of routines that survive unless the firm dies or manages to mutate its way of doing business. This latter capacity is limited by the limits of humans' ability to predict or foresee the future. There is, for writers such as Nelson, simply too much information to process: firms will usually lack the capacity to understand it all and understand what routines it needs to change to arrive at the best of all possible worlds. The odds, then, of any single actor treading the optimal path of technological development are exceedingly low.

When cognitive limitations combine with the phenomenon, in at least some markets, of path dependence (that is, technological "lock-in," or "network externalities"),[40] then reaching suboptimal technological outcomes is not only possible but likely. Evolutionists, pace Dilbert, consider firms to be unimaginative creatures whose ideas of the future tends to be closely tied to the present, like the 19th century farmer who asks for a better ox instead of a tractor. The "network" benefits of doing business in accord with the way everyone else does it adds to the problem. The result can quite easily become technological complacency, the graveyard of economic growth.

Here lies the link between neo-Schumpeterian economics and the e2e principle described in the opening section. The e2e principle can be understood as the implementation of an evolutionary innovation policy. E2e mandates that innovation is the job of the many (the ends), not the few (the center). By prescribing non-discrimination, it also sets conditions necessary for a fair fight, so that what survives is the truly the fittest and not merely the favored. E2e can help erase through

[38] SCHUMPETER, ECONOMIC DEVELOPMENT, *supra* note 29, at 93.

[39] *See, e.g.*, RICHARD NELSON & SIDNEY WINTER, AN EVOLUTIONARY THEORY OF ECONOMIC CHANGE 14 (1982) ("we reject the notion of maximizing behavior as an explanation of why decision rules are what they are").

[40] *See generally* W. BRIAN ARTHUR, INCREASING RETURNS AND PATH DEPENDENCY IN THE ECONOMY (1994).

competition the invariable mistakes that a centralized network planner will make.

This hostility toward centralized, planned innovation should, for Openists in particular, spill over to an attitude toward government. Government, no more than any human entity, is likely to have a good idea of what the future should be, so centralized technological planning is no better option. But the developments in evolutionary economics and post-Schumpeterian thought should direct Deregulationists to rethink their argument. It cannot be denied that the unregulated companies favored by the deregulation can become among the forces that resist the new. The new work suggests that this is not only possible, but likely.

All of these teachings lead to a single principle that should be an absolute policy consensus. Lost-cost market entry is the common foundation of the innovation theories that both Deregulationists and Openists subscribe to. That means preventing any single actor, governmental or otherwise, from becoming lord of the technological future. A multiplicity of innovating actors, even if suffering from the same inability to accurately predict the future, may nonetheless stumble upon the optimal path. But all should understand that the process will be an ugly, Darwinian affair, an interminable exercise in trial and error, and not the well-calculated elegance of monopolistic prophecy.

B. Vertical Integration & New Institutional Economics

While the study of vertical integration may seem to be a technical topic, it has become central to understanding the division between Openists and Deregulationists, and what the possibilities for reconciliation are.[41] For the work in this area proposes that the ends favored by Openists—namely, the innovations commons—may be reached by Deregulationist means. The analysis of vertical integration has highlighted weaknesses in the Openist position. Strong opposition to all vertical integration, expressed in the "open access" laws, has failed to answer to the challenge implicit in examples of "good" vertical coordination.

Why pay any attention to vertical integration at all? The specific reason is the "open access" debate. Some Openists, early on, suggested that the best means of preventing an erosion of the neutrality of the network would be laws limiting vertical integration of broadband carriers with Internet service providers. Keeping these two economic units separate, suggested Lawrence Lessig and Mark Lemley in early work, is

[41] A far better overview of this aspect of the debate is provided by Joseph Farrell & Philip J. Weiser*, Modularity, Vertical Integration, and Open Access Policies: Towards a Convergence of Antitrust and Regulation in the Internet Age,* 17 HARV. J.L. & TECH. 85 (2003).

likely to prevent content discrimination on the Internet.[42] The counter-argument is by now familiar for those who follow the debate. First, as Phil Weiser and Joseph Farrell reminded, vertical integration often leads to important efficiencies.[43] Second, as Jim Speta and others have pointed out, broadband operators, even if vertically integrated, want to make their product as valuable as possible and can therefore be expected to provide their customers with wide access to content and services.[44] Weiser and Farrell express this as the "ICE presumption," a presumption that a platform monopolist will "internalize complementary externalities."[45]

The literature has focused on a narrow but crucial question: how likely is it that private firms will create an innovation commons when that would be economically desirable? The answer begins by recognizing that the value of a broadband operator's (or any platform owner's) service ultimately depends on what applications and content it supports. The value of a game console to a consumer is chiefly a function of the games you can play on it (imagine an advanced game console that offered only "Pong"). We ought therefore expect the broadband operator to do everything possible to maximize the platform's value to its customers, including the adoption of whatever strategies will lead to the best environment for developing applications. For example, a service that only allowed Comcast customers to email Comcast customers would be less valuable, making it unlikely that Comcast would impose such a limitation. Similarly, if an "open" application development model yields the best applications, then the platform owner will provide an open model.

There may also be services where vertically coordinated, "hand-in-glove" cooperation results in more value for the customer. A car that arrived with no speedometer or tachometer would be less desirable despite the fact that the automobile and gauge market are arguably separate. In the broadband context, Comcast might, for example, want to offer its customers an integrated Voice-over-cable product. Doing so might be better with vertical coordination between itself and a telephony carrier. In short, some applications are better provided in a closed fashion, and some open. What is better open and better closed is

[42] *See* Mark A. Lemley & Lawrence Lessig, *The End of End-to-End: Preserving the Architecture of the Internet in the Broadband Era*, 48 UCLA L. REV. 925 (2001).

[43] See Farrell & Weiser, *supra* note 41, at 100-05.

[44] See James B. Speta, *A Vision of Openness by Government Fiat,* 96 NW. U. L. REV. 1553, 1565-66 (2001); James B. Speta, *Handicapping the Race for the Last Mile? A Critique of Open Access Rules for Broadband Platforms,* 17 YALE J. ON REG. 39 (2000).

[45] *See* Farrell & Weiser, *supra* note 41, at 101.

ultimately an empirical question,[46] and one that the platform owner—the argument goes—is best situated to answer.

But hold on: what if the platform owner is a monopolist—won't it try to "leverage" its platform monopoly into a second monopoly? For example, might a monopolist broadband operator begin to try and give itself a monopoly over all Voice-over-IP revenue? Here, for a Deregulationist, the relevance of the "single-monopoly profit" principles emerges. To a platform monopolist, the applications are its inputs, and the monopolist has the same interest as any other party in minimizing its input costs.[47] Hence, if allowing open application development saves the monopolist money, then it will do so. An example comes from Microsoft, monopoly owner of the Windows platform. Microsoft does not categorically bar any foreign applications, but instead integrates some functions into the operating system platform (such as, say, disk defragmentation utilities), and leaves others open to some degree of competition (such as word processors). While the merits of Microsoft as a model are debatable, the point is that even a monopoly platform owner may find it a bad idea to make everything vertically integrated.

This analysis leads to a presumption that, in the telecommunications market, vertically integrated companies, even with monopoly power, should generally be left unregulated, absent special conditions, or exceptions.[48]

But from both Weiser and Farrell's work, and from the evolutionary economic work discussed above, there is an important reason to suspect that platform owners may not implement optimal innovation policies themselves. Weiser and Farrell call it the problem of "incompetent incumbents."[49] In the terminology of Nelson and Winter, it is the observations that firms operate on the basis of routines that do not allow for suitable decisional flexibility. Perhaps most simply: the clearest problem is that no company will plan its own death, even if its death is in the social interest.

The problem for policy-makers is this: when a platform owner chooses a closed system, how can we know whether is it actually trying to "internalize complementary externalities" or just trying to protect itself? Is the platform owner truly creating a better product (like a car that

[46] *Cf.* Douglas Lichtman, *Property Rights in Emerging Platform Technologies,* 29 J. LEGAL STUDIES 615 (2000) (describing certain situations in which a platform owner might choose an open platform).

[47] *See* RICHARD A. POSNER, ANTITRUST LAW 177–78 (2d ed. 2001).

[48] Farrell and Weiser provide a useful summary of the exceptions that have emerged from the economic literature. Two are particularly relevant to the broadband context (1) interests in price discrimination and, (2) interests in disadvantaging potential platform rivals. *See* Farrell & Weiser, *supra* note 41, at 105-19.

[49] *Id.* at 114-17.

includes a speedometer) or is it, in Schumpeter's phraseology, "resisting to new ways" in an effort to prevent its own inevitable demise?[50] Effective competition will threaten the life of existing firms. As Schumpeter put it, in a true capitalist system, companies face "competition which commands a decisive cost or quality advantage and which strikes not at the margins of the profits and the outputs of the existing firms but at their foundations and their very lives."[51] If innovation presents a firm with a threat to its very existence, then its interest in a closed system may have much less to do with "internalization of complementary externalities" than it does with basic survival.

For policy-makers, the best answer to this dilemma, I believe, combines a program of education and regulatory threat. It is reasonable to agree that certain applications may be more efficiently provided open and others closed, and still see industry education as the primary challenge. Policy makers should be suspicious of the premise that internal processes of firm-decision will always or even often lead to good decisions. There are many reasons, not all of which can be discussed here, but one is that the information and signals that broadband operators are exposed to can be biased. Equipment vendors have spent years convincing broadband operators that great profits lie in capturing the applications market for themselves. In my personal experience, Wall Street analysts reward broadband operators in the short term for announcing plans to move into the applications market without serious analysis of the second-monopoly profit problem. Neither group has much to lose from sending such messages but both operators and consumers do. A vivid example came in 2000, when broadband operator Enron announced bold moves into the Video-on-Demand market and was cheered by financial and industry analysts (though obviously punished later).[52] In that case, the problem was not quite that the operator did not understand the one monopoly profit rule; it seemed to be that analysts did not seem to care.

This view sees industry education as paramount. One important tool in this respect is the regulatory threat, which can be important as a kind of signaling tool.[53] It can counteract information broadband operators get from other sources. Notably, FCC Chairman Michael Powell has taken steps toward such an educational policy. Powell has encouraged broadband owners to guarantee the neutrality of the network

[50] SCHUMPETER, ECONOMIC DEVELOPMENT, *supra* note 29, at 86.

[51] SCHUMPETER, CAPITALISM, SOCIALISM & DEMOCRACY, *supra* note 27, at 84.

[52] *See* Cecily Barnes, *Blockbuster Tests Video Streaming,* CNET NEWS.COM (Dec. 19, 2000) *at* http://news.com.com/2100-1023-250126.html.

[53] *See* Wu, *supra* note 2 (suggesting regulatory threat may force operators to consider the value of openness).

for their own sake as well as for that of consumers. His approach challenges operators to respect four "Internet freedoms" of the Internet consumer to guarantee a better network for all.[54] This message, if it reaches operators, may balance the urgings of others, such as equipment vendors and sometimes Wall Street, to seek a (unachievable) second monopoly profit.

III. RECONCILIATION

In what is perhaps an excess of optimism I consider reconciliation plausible. As the discussion above suggests, the insights of the Openists and Deregulationists are not necessarily in tension. Consider that both sides are basically interested in innovation and open market entry. The Openists are principally concerned with ends (an open network), and the Deregulationists, means (non-governmental methods). That suggests room for agreement.

A. Network Neutrality and the Model of Users' Rights

Based on the positions developed here, I believe neither Deregulationists nor Openists should oppose well-drafted Network Neutrality (NN) rules. Such NN rules are, ideally, users' rights to use the equipment or attachments that they want, following directly the open, deregulatory spirit of *Hush-A-Phone*. Neither side should have much reason to oppose a rule that creates a right of users to use whatever legal and non-harmful application "attachments" they want. NN rules, stated otherwise, can do much to advance the joint Schumpeterian interest in wide-open market entry.

NN rules are distinguished by creating rights in *users*. Rights, that is, to attach equipment or access any application or content, so long as it is not harmful or illegal. As a recent proposed rule reads:

> (b) *General Right of Unrestricted Network Usage.* Broadband Users have the right to use their Internet connection in ways which not unlawful or harmful to the network. Accordingly neither Broadband Operators nor the Federal Communications Commission shall impose restrictions on the use of an Internet connection except as necessary to: [prevent uses illegal under statute or uses harmful to the network].[55]

[54] *See* Michael K. Powell, *Preserving Internet Freedom: Guiding Principles For The Industry*, 3 J. ON TELECOMM. & HIGH TECH. L. 5 (2004).

[55] *See* Appendix A. This is the most recent version of regulations first proposed in an *ex parte* submission to the FCC by Tim Wu and Lawrence Lessig. *See* Tim Wu & Lawrence Lessig *Ex Parte* Letter, Appropriate Regulatory Treatment for Broadband Access to the

This distinguishes NN rules from competitor-centered rules like the various state-law "open access" regimes, or the approach of § 251 of the 1996 Telecommunications Act.[56] For example, the Portland merger condition at issue in the original AT&T open access case creates rights in ISPs, not users.[57]

The attraction to Openists of an NN rule is perhaps more intuitive. What is the attraction to Deregulationists? The key point is that creating rights in users can and will serve deregulatory purposes. American law is full of such deregulatory rights, economic and otherwise. A good example is the rights created by the dormant commerce clause to be free from discriminatory state regulation.[58] A user-centered NN rule is as deregulatory in spirit as *Hush-A-Phone* and *Carter-Phone*[59] were. It prevents government from acting as in the *Hush-A-Phone* case and agreeing to regulations that block application or network attachment. While less likely in recent years than in the 1950s and 1960s, the possibility of such action should not be discounted, for the reasons for doing so in the future cannot be predicted today. NN rules are, in short, like other rights against government: a form of pre-commitment rule for both government and industry. They prevent now what may be temptations tomorrow.

In addition, the broadband industry and some Deregulationists may be overlooking the extent to which NN rules prevent government from blocking *operator* entry into the application market. If the users have the right to access lawful applications and content, that includes those provided by the operator itself. NN rules prevent a quarantine—prevent operators from offering competitive, vertically integrated applications themselves. NN rules for these reasons have a value to the operator industry that should not be minimized.

Finally, NN rules are, at bottom, rules designed to free market entry, and should therefore be supported by those with Schumpeterian leanings, which means nearly everyone in communications policy. The NN rules create a structural bias that favors entry of any player, operator or application, or equipment-developer, into the market for consumer usage of the Internet. They are designed to make the Vonage story repeat itself. Even if Vonage dies, the Schumpeterian will admit it will have succeeded in bringing the network forward. The NN rules also do

Internet Over Cable Facilities, *Notice of Proposed Rulemaking*, FCC CS Docket No. 02-52 (filed Aug. 22, 2003), *available at* http://faculty.virginia.edu/timwu/wu_lessig_fcc.pdf.

[56] 47 U.S.C. § 252 (2000).

[57] AT&T Corp. v. City of Portland, 216 F.3d 871, 874 (9th Cir. 2000).

[58] *See, e.g.*, Kassel v. Consolidated Freightways Corp., 450 U.S. 662 (1981) (discussing the rights created by the dormant commerce clause).

[59] *See* Use of the Carterfone Device in Message Toll Tel. Serv., *Decision*, 13 F.C.C.2d 420 (1968).

not (as Christopher Yoo argues, discussed below) do anything in particular to prevent "facilities-based" entry. If anyone thinks they have a better idea than the TCP/IP protocol, they are free to build that network and see how it goes.

One Deregulationist who has not overlooked these arguments and the desirability of NN principles is FCC Chairman Michael Powell. Powell has spoken powerfully on the normative desirability of "Internet freedom," his phrase for network neutrality. "Internet freedom," he says, means "ensuring that consumers can obtain and use the content, applications and devices they want."[60] Doing so, he says, is "critical to unlocking the vast potential of the broadband Internet," and (in Schumpeterian language), "essential to nurturing competitive innovation."

Powell's discussion of "Internet freedom" focuses also on users' rights, notably, the four "freedoms" are:

> *Freedom to Access Content.* First, I believe consumers should have their choice of legal content.
>
> . . . .
>
> *Freedom to Use Applications.* Second, consumers should be able to run applications of their choice.
>
> . . . .
>
> *Freedom to Attach Personal Devices.* Third, consumers should be permitted to attach personal devices they choose to be the connections that they pay for in their homes.
>
> . . . .
>
> *Freedom to Obtain Service Plan Information.* Finally, and most importantly, consumers must receive clear and meaningful information regarding their service plans and what the limits of those plans are.[61]

These principles advocated by Powell, while done as part of an educational campaign, underline why Openists and Deregulationists should find common ground in advocacy in user-centered network neutrality rules. A shared faith in consumer choice and open market entry augurs such a result.

[60] Powell, *supra* note 54, at 12.

[61] *Id.*

B. *Criticism of Network Neutrality*

While some Deregulationists, like Chairman Powell, have endorsed principles of network neutrality, many industry players and some Deregulationists have mounted challenges to network neutrality proposals. I suggest that these challenges are generally lacking in merit, for reasons that follow.

The industry's most common challenge is this: while neutrality might be an attractive goal, any neutrality regulation is a solution looking for a problem. Such regulation or even a threat thereof, violates the principle of *Primum Non Nocere* (first, do no harm).[62] At its worst, network neutrality regulation might become a tool in the hands of application developers used to block competition from broadband operators. Imagine, for example, a rule that required FCC permission before a broadband operator could offer any service beyond a basic connection.

There are several problems with the *Primum Non Nocere* objection. First, it simply raises a question of dueling baselines. The existing design of the Internet is neutral. Why should it not be private entities who follow the principle of "do no harm" before monkeying with the proven strengths of the existing design? In this sense the slogan does nothing but restate an underlying difference in visions.

Second, the objection relies on an anti-regulatory straw-man. Because it is possible to imagine a *bad* network neutrality law, *any* network neutrality regulation is suspect. Yet it is unclear how Chairman Powell's or other's suggestions create the means for preventing competition among applications. The cable industry, the leading exponent of the do-no-harm view, has very meager support for its claim that a NN rule would block operator entry into the applications market. Its sole support is a proposal from Amazon that could be read to bar cable-operators from adding pop-up ads to web content.[63] That's far from a rule that prevents operators from entering the applications market. And as discussed above, a NN-rule that creates user's rights will give operators as much as anyone else a right to enter the applications or equipment markets.

A more powerful challenge to network neutrality rules runs as follows. It may be true that the basic, neutral Internet creates positive externalities, like the electrical grid or other neutral networks. But the metaphor is inapt for the following reason: the electric grid model fails to

[62] *See* Owen & Rosston, *supra* note 19.

[63] *See* National Cable and Telecommunications Association Ex Parte Letter, Appropriate Regulatory Treatment for Broadband Access to the Internet Over Cable Facilities, *Notice of Proposed Rulemaking,* FCC CS Docket No. 02-52 (filed Feb. 21, 2003).

take into account the possible need to improve the grid or infrastructure itself and the creation of proper incentives to do so. As Howard Shelanski puts the point, using roads as a metaphor: "at some point the road needs to be improved and that work can be disruptive. So the question is not one of never disrupting the flow of traffic, but of knowing when to let cars run freely on the road and when to tear up the road to make it work better."[64]

This returns us to the "smart-pipe" discussion and the argument that much innovative potential is trapped in the core of the network, a point Christopher Yoo makes.[65] Yoo argues that it is critical, in a market with many vertical layers, that competition be encouraged at the layer that is least competitive. As he states, "Application of the basic insights of vertical integration theory reveals that markets will achieve economic efficiency only if each stage of production is competitive."[66] Looking at broadband, he thinks that in the application and content market, competition is robust and needs no favors. Yet he sees competition at the physical layer (between cable and DSL) least vigorous and therefore the most in need of freedom from government restraints. Network neutrality regulation, in Yoo's view, would mandate dumbness and therefore slow deployment of proprietary "smart" networks.[67]

According to Yoo, the answer is to allow or even encourage the deployment of divergent proprietary, as opposed to standardized, broadband networks. He sketches the possibility of consumers being served by three entirely different and non-standardized broadband infrastructures: "The first network could be optimized for conventional Internet applications, such as e-mail and website access. The second network could incorporate security features designed to appeal to users focusing on e-commerce. The third network could prioritize packet routing in the manner needed to facilitate time-sensitive applications such as VoIP."

Yoo's conclusions are overstated and demand several responses. First, it is unclear why Yoo believes that the existence of a neutral Internet would be a barrier to "facilities-based competition," that is, the market entry of entirely new network facilities.[68] If an operator wanted

[64] Howard Shelanski, Remarks at Silicon Flatirons Telecommunications Program Conference, University of Colorado School of Law (Feb. 8, 2004) (transcript available from the Silicon Flatirons Telecommunications Program, http://www.silicon-flatirons.org).

[65] Yoo, *supra* note 8, at 42-46.

[66] *Id.* at 59.

[67] Adam D. Thierer makes the same point. *See* Adam D. Thierer, *Are 'Dumb Pipe' Mandates Smart Public Policy? Vertical Integration, 'Net Neutrality,' and the Network Layers Model*, 3 J. ON TELECOMM. & HIGH TECH. L. (forthcoming Winter 2005).

[68] *Cf.* Randal Picker, *Entry, Access and Facilities-Based Competition*, *in* AM. L. & ECON. ASS'N ANN. MEETINGS (The Berkeley Electronic Press Working Paper No. 33, Apr. 29, 2004).

to build an entirely new network designed, say, to offer voice services, it is free to do so. The existence of the Internet for new facilities deployment seems irrelevant. Indeed, Yoo seems to have it backward: if the neutral network is no good for certain applications, that would drive facilities-based competition, not inhibit it. A neutral network should be expected to drive an efficient mix of shared and facilities-based competition: those applications which can be run over the open network will be, and for those that require entirely new facilities, new facilities will be built. Much of the cell-phone networks, for example, were built in the 1990s, and the Internet proved no barrier.

In fact the facilities-based competition that Yoo sees as ideal is our present reality. The existing telephone network is Yoo's "prioritized" network that facilitates a time-sensitive application, telephony, as are the mobile-phone networks. Meanwhile, the cable television network is a network specialized for "one-to-many" video. Perhaps Yoo's point is that these various specialized networks are likely to remain in our lives, but that doesn't say much about how the Internet should be regulated.

Second, Yoo's premise that vigorous competition at every layer is always better for the consumer is overstated. He downplays, to the point of elimination, the basic economic benefits of standardization. And when it comes to technology platforms or other areas of economic development it is easy to envision scenarios where standardization means less competition but is nonetheless socially beneficial, which impeaches Yoo's premise.

Here is an intuitive demonstration of the point. Most people in the United States speak a standard language, English. This undoubtedly leads to some sacrifice. We lose, for example, the precision of German; we lack the Chinese vocabulary for food; and we lose righteousness and occasional elegance of the French language. But few would argue that vigorous and ongoing competition for a standard American language would clearly serve consumer welfare. It would be, instead, the Tower of Babel.

The same observation holds for standardized technology platforms such as the Windows operating system or the TCP/IP protocol, which bring a variety of benefits for application developers and end users. Application writers need only write for a single platform, for example, and can expect to reach a much larger addressable market, thereby justifying greater investments. End-users, given a single standard, share information with ease. All of these advantages usually go under the rough heading of network externalities, or the economic benefits of standardization. Yoo is, in essence, failing to take seriously the benefits of platform standardization in his product differentiation model. To be sure, as with language, there are costs from uncompetitive platform markets. The result will in all likelihood be an inferior platform (for want of competition), and the possibility of anti-competitive conduct. But the fact that we face a balance of costs and benefits shrinks Yoo's point. We are left instead with the empirical question: how valuable are neutral standards and networks, and when are they worth a loss in competition in the network?

Yoo and others who favor the encouragement of market entry should in fact favor basic network neutrality rules. True enough, such rules may slow some competition for the standards for the Internet's basic protocols. But if that's truly the case, nothing in NN rules, prevent full facilities-based competition. And meanwhile NN rules facilitate market entry on the standardized and highly successful network we do have. These and other reasons should prompt those Deregulationists opposed to network neutrality principles to ask whether they are on the wrong side of the argument.

CONCLUSION

I've suggested here that reconciliation of the broadband debate is plausible, but unfortunately that doesn't make it inevitable. A serious contribution to this problem has come from the winner-take-all approach of some of the groups on each side. The Internet Service Providers have seemed committed to achieving full open access rules through litigation, again showing that companies in fear of death turn to lawyers with the same urgency that dying people turn to doctors. And the cable industry, while it has laudably adhered to neutral practices during the last period of intense scrutiny, still seems unwilling to agree with a simple neutrality rule that would codify its existing practices and do much to remove regulatory scrutiny. As this goes, it should be recognized that the age of regulatory uncertainty surrounding broadband will soon reach its first decade. That fact alone should prompt all interested parties to seek reconciliation sooner rather than later.

APPENDIX A:
DRAFT NETWORK NEUTRALITY RULE

§ 1. *General Right to Unrestricted Network Usage.* Broadband Users have the right reasonably to use their Internet connection in ways which are not illegal or harmful to the network. Accordingly neither Broadband Operators nor the Federal Communications Commission shall impose restrictions on the use of an Internet connection except as necessary to:

(1) Comply with any legal duty created by federal, state or local statute, or as necessary to comply with any executive order, warrant, legal injunction, subpoena, or other duly authorized governmental directive;

(2) Prevent physical harm to the local Broadband Network caused by any network attachment or network usage;

(3) Prevent Broadband users from interfering with other Broadband or Internet Users' use of their Internet connections, including but not limited to neutral limits on bandwidth usage, limits on mass transmission of unsolicited email, and limits on the distribution of computer viruses, worms, and limits on denial-of service-or other attacks on others;

(4) Prevent violations of the security of the Broadband network, including all efforts to gain unauthorized access to computers on the Broadband network or Internet;

(5) Serve any other purpose specifically authorized by the Federal Communications Commission, based on a weighing of the specific costs and benefit of the restriction.

§ 2. As used in this section,

(1) "Broadband Operators" means a service provider that provides high-speed connections to the Internet using whatever technology, including but not limited to cable networks, telephone networks, fiber optic connections, and wireless transmission;

(2) "Broadband Users" means residential and business customers of a Broadband Operator;

(3) "Broadband Network" means the physical network owned and operated by the Broadband Operator;

(4) "Restrictions on the Use of an Internet Connection" means any contractual, technical, or other limits placed with or without notice on the Broadband user's Internet Connection.

COMPETITION POLICY FOR MOBILE BROADBAND NETWORKS

HOWARD A. SHELANSKI[*]

TABLE OF CONTENTS

INTRODUCTION

Over the past decade, mobile wireless systems have changed from analog cellular technology to digital networks and have more recently been moving to higher capacity networks capable of supporting broadband services. One commentary refers to mobile broadband services as "melding two popular innovations: the Internet and mobile technologies."[1] High-speed mobile services are often referred to as "3G"

[*] Professor of Law, University of California at Berkeley. This essay is based on a presentation I made as a panelist at the Silicon Flatirons Symposium on "Digital Broadband Migration: Toward a Regulatory Regime for the Internet Age," February 8, 2004, at The University of Colorado.

[1] MARTIN BAILY, ET AL., CELLULAR TELECOMMS. & INTERNET ASS'N, AN ECONOMIC ANALYSIS OF SPECTRUM ALLOCATION AND ADVANCED WIRELESS SERVICES (Oct. 2001), *available at* http://www.sbgo.com/Papers/An%20Economic%20Analysis%20of%20Spectrum%20Allocation.pdf.

(for "3rd Generation") or now even "4G" services.[2] The Federal Communications Commission (FCC or Commission) defines as "3G" those mobile services that can support data transport rates of at least 144 kilobits per second and up to 2 megabits per second; that are provided over systems with a high degree of global compatibility and interoperability; and that can support a wide range of voice and data applications.[3] 3G capabilities are thus comparable to current mass-market broadband technologies, like Digital Subscriber Line (DSL) and cable modem service, that meet the FCC's definition of high-speed communications.[4] The race to build such high-speed mobile networks is being driven by the increasing volume of wireless data traffic, which, according to estimates in one FCC report, may overtake the volume of wireless voice traffic by 2006.[5]

The development of mobile broadband technology—and of wireless Internet networks generally—has implications for a variety of current issues in telecommunications policy. One particularly interesting set of issues involves variations on the question of whether regulators should require systems to be open to all users or, conversely, whether regulators should allow proprietary systems to exclude or discriminate against access by others. For example, must Internet transport networks provide a neutral, "end-to-end" conduit for all content and services or may they favor some content/service providers over others that traverse their networks to consumers? To what extent should the FCC make radio spectrum a commons open for use by all (subject to non-interference) and to what extent should it license frequencies for exclusive use and control by specified users? Should network-operating standards be open and common or should they be proprietary and competitive? The emergence of mobile broadband networks affects, and in turn will likely be affected by, the answer to each of the above questions.

Consider first the question of end-to-end requirements for Internet transport. Whether, and to what extent, owners of networks that carry Internet traffic to consumers must make their networks open on a non-discriminatory basis to content/service providers has become a hotly debated question. On one side, commentators argue that absent such

[2] *Move over 3G: here comes 4G*, ECONOMIST, May 29, 2003, *available at* http://www.economist.com/business/displayStory.cfm?story_id=1816742.

[3] FCC, THIRD GENERATION WIRELESS SYSTEMS, *at* http://www.fcc.gov/3G/ (last updated Nov. 25, 2002).

[4] Inquiry Concerning the Deployment of Advanced Telecomms. Capability to All Americans in a Reasonable & Timely Fashion, & Possible Steps to Accelerate Such Deployment Pursuant to § 706 of the Telecomms. Act of 1996, *Notice of Inquiry*, 19 FCC Rcd. 5136, 5139-40 ¶ 11 (2004).

[5] Service Rules for Advanced Wireless Services in the 1.7GHz & 2.1 GHz Bands, *Notice of Proposed Rulemaking*, 17 FCC Rcd. 24,135, 24,138 ¶ 6 (2002).

end-to-end openness, network owners will extend their control from the transport layer to the applications layer thereby deterring the innovation that has brought consumers the enormous range of content and services they can now receive on line.[6] On the other side are commentators that argue such regulation is unnecessary given the economic incentives of competing networks and, moreover, that such regulation could interfere with positive vertical relationships that enhance innovation and benefit consumers.[7] While the end-to-end principle has important virtues, the principle's benefits will under certain conditions come with offsetting costs for consumer welfare and network innovation.

Wireless broadband is relevant to the end-to-end debate because its development will directly affect the question raised above: where should policy makers draw the line between end-to-end mandates and the potential benefits of proprietary network innovation and of vertical relationships between transport platforms and content/services? The concern about vertical discrimination by network owners in favor of some content/service providers and against others is made particularly acute by the paucity of broadband alternatives to which consumers currently have access. Most consumers can currently choose from at best two options: DSL over the local telephone network and cable-modem service over the local cable system. In such a concentrated market, content and service providers that are not favored by the DSL or cable-modem provider might have difficulty reaching consumers and gaining a foothold in the market. The more networks there are, however, the greater the opportunity for content/service providers to gain high-quality transport and the greater the ability of consumers to vote with their dollars for the content/services they want by choosing to subscribe to different systems. Thus, the growth of wireless broadband increases competition among networks and expands consumer choice, diminishing the case for mandatory, end-to-end openness.

Similarly, because mobile broadband will require consistent access to substantial amounts of spectrum but could also attract new entrants and technologies, it raises important questions about the balance between licensing and commons approaches to spectrum assignment.[8] In the

[6] *See* Mark A. Lemley & Lawrence Lessig, *The End of End-to-End: Preserving the Architecture of the Internet in the Broadband Era*, 48 UCLA L. REV. 925 (2001); Tim Wu, *Network Neutrality, Broadband Discrimination*, 2 J. ON TELECOMM. & HIGH TECH. L. 141 (2003).

[7] *See* James B. Speta, *The Vertical Dimension of Cable Open Access*, 71 U. COLO. L. REV. 975 (2000); Philip J. Weiser, *The Internet, Innovation, and Intellectual Property Policy*, 103 COLUM. L. REV. 534 (2003); Thomas J. Tauke, *Current Regulatory Realities: Overcoming the Regulatory Quandary*, 3 MICH. ST. DCL L. REV. 609 (2003).

[8] *See* Kevin Werbach, *Supercommons: Toward a Unified Theory of Wireless Communications*, 82 TEX. L. REV. 863 (2004); Stuart Minor Benjamin, *Spectrum*

United States, all radio spectrum is legally the property of the U.S. government.[9] The government then decides which frequencies will be available for non-government uses and the FCC allocates that spectrum for particular uses (e.g., TV, FM radio, and wireless telephone service) and assigns it to particular users through its licensing process. Those licensees in turn have broad and renewable property rights in their assigned frequencies that enable them to exclude other users from their spectrum, even if those other users would not interfere with the licensee's transmissions.[10] The more spectrum one has available, the more information can be transmitted. Incumbent wireless operators upgrading their networks to achieve broadband capabilities will thus want to preserve and expand their proprietary access rights to choice frequencies to accommodate their higher capacity systems.

New entrants into the wireless marketplace may, however, call for a commons or some other access regime under which the incumbent property rights cannot block them from operating in a non-interfering manner. Can the technologies, like spread-spectrum, that today permit simultaneous use of the same frequencies, scale to the capacity demands of wireless broadband? If not, will giving priority to certain users deter innovation by others that would allow a commons approach for mobile broadband? Again, the development of a competitive mobile broadband market will be essential to assuring that the potentially adverse consequences of entrenched spectrum rights are mitigated and that mobile broadband markets deliver both the short-run benefits of price competition and the long run benefits of innovation to consumers.

Finally, 3G raises the question of whether to have competing or common standards, an important decision for any emerging network technology. In mobile services, in particular, there has been much debate over whether the industry has developed better in Europe where there is a common GSM standard or in the United States where carriers compete as much on their underlying technologies as on their services.[11] As I will discuss further in this paper, the standards debate will be integral to the development of competition policy for mobile broadband services as well.

As the foregoing discussion makes clear, the development of mobile broadband networks raises a number of technological, economic, and legal questions. One challenge that lies at the intersection of those three

Abundance and the Choice Between Private and Public Control, 78 N.Y.U. L. REV. 2007 (2003).

[9] 47 U.S.C. § 301 (2000).

[10] *See* Howard A. Shelanski & Peter W. Huber, *Administrative Creation of Property Rights to Radio Spectrum*, 41 J.L. & ECON. 581 (1998).

[11] *See* Neil Gandal et al., *Standards in Wireless Telephone Networks*, 27 TELECOMM. POL'Y 325 (2003) [hereinafter *Standards in Wireless Telephone Networks*].

forces is the design of a framework for evaluating and protecting competitive performance of the mobile broadband market. Indeed, wireless broadband demonstrates how few are the degrees of separation between analogous but seemingly disconnected debates over end-to-end Internet transport rules, spectrum assignment policies, and standard setting in telecommunications. Mobile broadband services will not make those questions moot, but the development of a healthy and competitive mobile broadband market will affect where the policy cuts should be made and how, in turn, alternative policy decisions will affect consumers.

The remainder of this paper will thus discuss what, looking forward, is the appropriate competition policy framework for the mobile broadband industry. The answer will depend as a preliminary matter on which objectives policy makers choose to pursue. The various debates over deployment of advanced wireless services raise several, potentially inconsistent, goals that might affect a government's choice of antitrust regime for the industry. Consider just the following possible objectives: national leadership in the world market for wireless services; a highly competitive domestic market to maximize long-run economic benefits to subscribers; speeding deployment of mobile broadband networks; or, ensuring the development and deployment of the best possible technology for mobile broadband networks. That these goals would co-exist uneasily is evident. For example, if speed of deployment is paramount, then measures to facilitate rapid construction of networks using today's most quickly deployable technology should be taken. Yet such measures run the risk of locking in, for a period of time, a technology that is not the best one currently or imminently available. If a country deems global leadership in the sector to be a priority, then collaboration among domestic service providers might be tolerated notwithstanding its impact on domestic competition. The point, in brief, is that optimal policy depends on what one wants to maximize.

The discussion that follows will assume that the objective of competition policy for the mobile wireless Internet industry is to maximize long-run consumer welfare, which is essentially the objective of modern antitrust (or competition) law in the United States, the European Union, and increasingly in other jurisdictions.[12] The selection

[12] *See, e.g.*, U.S. DEPARTMENT OF JUSTICE & FEDERAL TRADE COMMISSION, 1992 HORIZONTAL MERGER GUIDELINES, 57 Fed. Reg. 41,552 (1992), *revised*, 4 Trade Reg. Rep. (CCH) ¶ 13,104 (Apr. 8, 1997), *available at* http://www.usdoj.gov/atr/public/guidelines/horiz_book/hmg1.html [hereinafter HORIZONTAL MERGER GUIDELINES]; EUROPEAN COMMISSION DIRECTORATE-GENERAL FOR COMPETITION, EUROPEAN COMMUNITY COMPETITION POLICY, XXXTH REPORT ON COMPETITION POLICY ¶ 1 (2001); ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, COMPETITION LAW AND POLICY IN MEXICO: AN OECD PEER REVIEW (2004), *available at* http://www.oecd.org/ dataoecd/57/9/31430869.pdf.

of a competition policy objective does not, however, lead inexorably to a clear and specific set of policies themselves, particularly in an emerging network industry like wireless Internet services. To see why, consider first the factors that a welfare-maximizing competition policy must encompass under existing general antitrust frameworks: (1) proper definition of the relevant market; (2) analysis of industry-specific barriers to entry; (3) determination of whether standards competition or cooperative standard-setting should be pursued; and (4) assessment of whether fostering innovation in the particular industry at issue has implications for market structure that differ from the structural assumptions for promoting short-run efficiency of prices and output in the relevant market. In addition, the administrative question of what kinds of institutions—e.g., general competition authorities or sector specific regulators—should be responsible for enforcing and implementing the policies must be decided.

This paper will discuss each of the above questions in the context of mobile broadband services. The purpose of this discussion is not to present an exhaustive or definitive set of policy prescriptions but instead to describe the central dimensions of competition policy for the mobile broadband market, to examine important and distinguishing features of the industry that affect the applicable antitrust regime, to analyze the tradeoffs among feasible policy choices and, finally, to present the important features and institutional framework that competition policy for the mobile broadband industry should incorporate. Section I of this paper will examine key aspects of competition policy for the mobile broadband market. Part A will discuss how to define the relevant market for 3G services. Part B addresses the benchmark for deciding whether the mobile broadband market, once defined, is "competitive." Part C discusses dynamic competition and possible tradeoffs with innovation. Part D turns to the issue of cooperative versus competitive standard setting, while Part E addresses the related question of interconnection among competing wireless networks. Section II of the paper will turn briefly to the institutional question of whether antitrust agencies or sector-specific telecommunications regulators should have the leading role in setting and enforcing competition policy in the mobile broadband market.

I. CENTRAL DIMENSIONS OF COMPETITION POLICY FOR MOBILE BROADBAND SERVICES

This section will address four important dimensions of competition policy for an evolving network industry and discuss how they apply to mobile broadband services. It will first address the conventional questions of market definition and competitive benchmarking for mobile broadband services. It will then address three issues particularly relevant to the dynamic technological environment of wireless Internet, which are the questions of tradeoffs between competition and innovation, of standard setting in the advancing wireless marketplace, and of interconnection among competing networks.

A. Market Definition for Advanced Mobile Services

In designing competition policy for an industry, the first step conventionally is to define the relevant market(s) in which that industry operates, in order to determine market structure and assess the prospects for exercise of market power. A long-standing principle by which economists define the product scope of a market is to include two goods or services in the same relevant market if consumers view them as sufficiently close substitutes and not to include them in the same relevant market if consumers do not view them as sufficiently close substitutes.[13] A similar logic is used for geographic scope. When are substitutes "sufficiently" close to being included in the same market? To some extent, toothpaste competes with clothing for consumers' dollars, but one should not conclude that toothpaste and clothes are in the same product market. To give more precision to the concept of sufficiently close substitutes, economists undertaking market delineation exercises often conduct a so-called hypothetical monopolist test. This test asks whether a hypothetical, profit-maximizing monopolist over a group of products in a given area could profitably raise prices above a specified level by a "small but significant" amount for a sustained period of time.[14] The group of products considered in the test is a candidate relevant market. The smallest group of products that satisfies the test constitutes a relevant market.[15]

[13] *See* George W. Stocking & William F. Mueller, *The Cellophane Case and the New Competition*, 45 AM. ECON. REV. 29, 44-48 (1955).

[14] HORIZONTAL MERGER GUIDELINES, *supra* note 12, at § 1.0; FTC v. Swedish Match, 131 F. Supp. 2d 151, 160 (D.D.C. 2000); California v. Sutter Health Sys., 130 F. Supp. 2d 1109, 1120 (N.D. Cal. 2001). *See also* Michael L. Katz & Carl Shapiro, *Critical Loss: Let's Tell the Whole Story*, 17 ANTITRUST 49, 53 (2003).

[15] HORIZONTAL MERGER GUIDELINES, *supra* note 12, at §§ 1.0, 1.11.

A price increase will raise a hypothetical monopolist's profits unless unit sales volume falls sufficiently to offset the higher price received for the units sold. Thus, the hypothetical monopolist test indicates that a set of products constitutes a relevant market if the hypothetical monopolist could make a "small but significant and non-transitory" increase in price without causing enough consumers to switch to substitute goods that the price increase becomes unprofitable.[16]

So what is the product or service that mobile broadband operators will compete to provide? Third and subsequent generation wireless networks will provide voice telephony but, more importantly, high-speed data services. If one were to define the market as "mobile voice and high-speed data" services, then the relevant market structure would depend only on the number of mobile broadband networks operating in the relevant geographical territory. A difficult initial question for mobile broadband market definition is, however, whether the market definition should be limited to mobile services or include other wireless services (e.g., WiFi), or be expanded still further to include wireline voice and broadband telecommunications services.

If mobile broadband services meet performance expectations, they will provide direct competition to wireline services like cable modem and DSL connections.[17] This does not mean that fixed and mobile broadband services should always be considered to be in the same market, however. The reciprocal competitive effect of fixed services on mobile wireless services need not be symmetric, and in fact is unlikely to be. For, to the extent that mobility has value to consumers, wireline voice and broadband services will not substitute for mobile wireless services.

Although the existence of fixed, wireline access technologies certainly creates some competitive pressure and pricing discipline for prospective mobile broadband service providers, there are several reasons why competition policy makers might not define the mobile broadband market to include wireline service providers. First, as mentioned, mobility itself has value to consumers. There is casual yet strong evidence of this proposition in the fact that most subscribers to wireless telephony in the United States also have landline telephone service.[18] Thus, all other features (e.g., speed, quality, reliability) equal, mobile

[16] *Id.* at § 1.0.

[17] *See* Jerry A. Hausman, *From 2G to 3G: Wireless Competition in Internet Related Services, in* BROADBAND: SHOULD WE REGULATE HIGH SPEED INTERNET ACCESS? 119-20 (Robert W. Crandall & James Alleman eds., 2002).

[18] FCC data show that 95.1% of U.S. households subscribe to conventional local telephone service. Press Release, FCC, Federal Communications Commission Releases Study on Telephone Trends (May, 2004), *available at* http://www.fcc.gov/Bureaus/Common_Carrier/Reports/FCC-State_Link/IAD/trend504.pdf.

broadband services would have an intrinsic advantage over wireline services that would enable mobile broadband operators to raise prices on their service without losing material numbers of customers to providers of fixed broadband services. Second, mobile and wireline broadband options might be imperfect substitutes because their distinct comparative advantages may lead them to be used for differing sets of applications in ways that limit substitutability. For example, mobile broadband services might be quite useful for businesses that involve employees in the field who have particular data and applications needs—for example the ability to relay and process order information quickly, to provide confirmation of product inventory, or to fill service orders from remote locations. Hardware and software tailored to such applications might be developed to work over mobile broadband networks but not for fixed broadband technologies. Consequently, even if mobile broadband services reach sufficient speed and reliability to substitute for wireline broadband, the reverse may not hold when antitrust issues related to mobile broadband competition are at issue.

A related question is whether less advanced forms of mobile services—i.e., narrowband PCS services—should be included in the mobile broadband market. This question may be harder to answer. On one hand, much will depend on the purposes for which consumers actually use mobile broadband networks. If consumers use mobile broadband mostly for voice and simple text messaging, then 2G networks might provide some level of substitution. A stronger reason for including 2G services in the relevant market, however, is that those networks are likely entrants into mobile broadband services. One of the accepted mobile broadband standards (the EDGE standard)[19] is in fact geared specifically to transitioning 2G TDMA networks to mobile broadband capability while the dominant 3G standard in the United States, CDMA2000, is designed for easy transition of CDMA-based 2G systems to 3G capabilities.[20] Because 2G networks might therefore become sources of supply elasticity that limit the market power of any mobile broadband networks, there is a good argument for adopting a dynamic perspective and including 2G networks in the mobile broadband market. But in the end, a careful analysis of subscriber switching costs and of the timeline for 2G conversion will have to be undertaken to make a conclusive judgment about whether the mobile broadband market should be defined to include remaining 2G networks. A weaker initial presumption might attach to restricting the market definition to existing or imminent mobile broadband providers and

[19] *See infra* note 44 and accompanying text.

[20] *Standards in Wireless Telephone Networks*, *supra* note 11, at 325.

excluding 2G networks. That presumption should be rebuttable by evidence that 2G substitutes for mobile broadband services or that 2G networks could convert to mobile broadband within a reasonably short time frame.[21]

B. Defining "Competitive" in the Context of the Mobile Broadband Market

Once the market definition exercise discussed above is completed, the next step in the competitive analysis is to consider what, given the particular technological and economic characteristics of mobile broadband service, would constitute a "competitive" market. How many mobile broadband networks can potentially enter the market? What barriers to entry are likely to arise for new entrants? In this regard, the most salient aspect of mobile broadband is its need for spectrum to be allocated for the service.

At present, there are about 180 MHz of conventional commercial mobile radio service (CMRS) spectrum available to provide mobile telephone service in each geographical market nationwide. In addition to this spectrum, the FCC has been working to auction an additional 78 MHz of spectrum in the 700 MHz UHF bands and 30 MHz of spectrum in the 2GHz satellite bands, which would be available for mobile broadband providers among others.[22] The Commission has also, working in conjunction with the National Telecommunications and Information Administration (NTIA), allocated (but not yet assigned to users) an additional 90 MHz of spectrum in the 1710-1755 MHz and 2110-2155 MHz bands specifically for mobile broadband use.[23] Other efforts to increase available spectrum are also underway at the FCC.[24] The Commission's attention to increasing available spectrum for mobile broadband has been in response to Congress's mandate that an additional 200 MHz of spectrum be made available for advanced wireless telecommunications.[25]

Assuming existing CMRS spectrum, over which consumers now receive wireless telephone service, can be "re-harvested" for mobile broadband purposes and adding the prospective 200 MHz of new spectrum, a total of roughly 400 MHz may be available for mobile

[21] HORIZONTAL MERGER GUIDELINES, *supra* note 12, at § 3.2 (defining an entry that could occur within 2 years as "timely" and competitively significant).

[22] Implementation of § 6002(B) of the Omnibus Reconciliation Act of 1993, *Eighth Report*, 18 FCC Rcd. 14,783, ¶¶ 26, 31 (2003).

[23] *Id.* at ¶ 31.

[24] *Id.* at ¶ 32.

[25] Omnibus Budget Reconciliation Act of 1993 § 113(b)(1), 47 U.S.C. § 923(b)(1) (2000).

broadband and other advanced wireless services in the next few years. Although it is unclear how much spectrum a mobile broadband operator needs to provide service, the planned spectrum allocation could support a number of rival providers. The market for mobile broadband services therefore has potential to be competitive, although the substantial fixed costs of providing the services suggest the market will not approach the idealized competition among atomized, price-taking firms found in textbooks.

Economic factors like network externalities or increasing returns to scale might further limit the number of competing networks notwithstanding the number that the above discussion suggests is technologically feasible. If, for example, consumers for some reason could obtain certain services on one network but not others, or if one network could serve all users at a lower per-subscriber cost than could multiple networks, then monopoly might develop and even have theoretical benefits. But interconnection among wireless networks (to be discussed further, below) will prevent any system from closing itself to calls originating on competing systems, thus eliminating "network externalities" that could lead to monopoly. Moreover, there is no evidence that mobile broadband networks will have cost structures that approach natural monopoly or that, in the end, will be substantially different in shape from the cost curves for conventional wireless networks now in place.[26] To be sure, there will likely be economies of scale over a certain range of demand. Any time a firm incurs the high, up-front, fixed costs of building a network, the average cost of serving each customer will decline for some time with each new network user. The economic limits on the number of firms the mobile broadband market can efficiently support will thus depend on the ultimate market demand for mobile broadband services and the number of efficient-scale firms that such demand can support. To the extent that the feasible number of efficient firms is smaller than the number of licenses the FCC allocates, consolidation will occur in the mobile broadband industry. Before presuming against mergers among mobile broadband providers, competition officials should take account of scale efficiencies and be careful to adopt a realistic benchmark for competition in the industry.

The above discussion is not intended to suggest that competition policy should, *ex ante*, target any particular number of firms as desirable in the mobile broadband market. Nor is it meant to cast doubt on the viability of competition among providers of mobile broadband services. Indeed, the analysis presumes sufficient competitiveness in the market

[26] For a general discussion of natural monopoly conditions, see STUART MINOR BENJAMIN ET AL., TELECOMMUNICATIONS LAW AND POLICY 374 (2001).

that general antitrust principles are likely to apply meaningfully in the mobile broadband marketplace. In the past, the Commission has prejudged the minimum, acceptable level of competition in wireless telecommunications. The Commission imposed a "spectrum cap" that prohibited any single firm from holding licenses to more than 45 MHz of the 180 MHz of CMRS spectrum available in a given geographical market, thus assuring the existence of at least four competitors. The Commission in 2001 eliminated the cap effective in 2003, and raised the cap to 55 MHz in the interim.[27] Part of the motivation for lifting the cap was concern that it artificially constrained firms from obtaining the spectrum they might find necessary for mobile broadband services, and thereby might deter investment in developing mobile broadband networks.

Although antitrust policies such as the U.S. Department of Justice/Federal Trade Commission (FTC) Horizontal Merger Guidelines and the European Union's Guidelines on the Assessment of Horizontal Mergers provide no rigid limits on concentration like those the spectrum cap imposed, they do provide useful presumptive limits on acceptable changes in market concentration through merger and acquisition. Application of those guidelines always depends to some extent on the specific market context and specific industry factors. In an evolving network industry like mobile broadband communications, this more flexible approach of antitrust policy has advantages over the categorical limits of rules like the spectrum cap because the benchmarks for assessing market performance can be more easily adjusted as the industry develops and competition authorities learn more about the economics of the relevant market.

C. Innovation and Competition in Mobile Broadband: Assessing Claims of Dynamic Tradeoffs

Related to the above discussion of establishing the right benchmarks against which to assess economic performance of the mobile broadband market is the question of the relationship between static and dynamic market performance. Participants in regulatory and antitrust proceedings affecting telecommunications have, with increasing frequency, asserted that policy decisions designed to promote or preserve competition will have unintended, negative consequences for technological change. The perceived role of technological change in the growth of the U.S. economy during the 1990's caused policy makers and consumers alike to pay greater attention to how innovation can increase economic welfare.

[27] 2000 Biennial Regulatory Review Spectrum Aggregation Limits For Commercial Mobile Radio Services, *Report & Order*, 16 FCC Rcd. 22,668 (2001).

One manifestation of this attention to innovation has been heightened sensitivity to whether the goals or presumptions of existing public policies might conflict with the goal of technological progress.[28] Whether regulators must sometimes make tradeoffs between innovation tomorrow and efficient resource allocation today has been debated in such diverse contexts as environmental regulation and antitrust policy.[29] The ways in which antitrust law might affect cooperative approaches to innovation has been an area of intense inquiry in recent years.[30]

The question of how policy affects technological innovation is especially salient in the telecommunications sector. Several kinds of policy arguments hinge on innovation. The most common form of the argument, made by participants in recent proceedings at the FCC and the Department of Justice, is that innovation may suffer if regulators focus too narrowly on preserving or improving competition in existing markets. The debate that surrounded the spectrum cap is a good example. In the FCC's 1999 proceedings on whether to retain the 45 MHz cap,[31] several carriers argued that consolidation of competing licenses was a necessary condition for the development of mobile broadband services.[32] Those carriers argued that without consolidation, they would be uncertain of having sufficient spectrum capacity for the new services and hence would find it too risky to invest in developing the new technology. As another example, in the FCC's 1999 rulemaking proceeding that limited the number of subscribers a single cable company could serve, some cable operators similarly argued that the introduction of broadband and telephone services on cable networks requires large-scale systems.[33]

The Commission addressed the above challenges in a case-by-case manner and, each time, at least initially maintained its emphasis on competition and static efficiency. In the 1999 spectrum cap proceeding, the Commission retained the 45 MHz limit in the interests of preserving

[28] *The Annual Report of the Council of Economic Advisors*, *in* ECONOMIC REPORT OF THE PRESIDENT 173-93 (1999) *available at* http://www.gpoaccess.gov/usbudget/fy00/pdf/1999_erp.pdf [hereinafter ECONOMIC REPORT OF THE PRESIDENT].

[29] *See id.*

[30] *See, e.g.*, Christopher Pleatsokis & David Teece, *The Analysis of Market Definition and Market Power in the Context of Rapid Innovation*, 19 INT'L J. INDUST. ORG. 665 (2001); David B. Audretsch et al, *Competition Policy in Dynamic Markets*, 19 INT'L J. INDUST.ORG.613 (2001); ANTITRUST, INNOVATION AND COMPETITIVENESS (Thomas M. Jorde & David J. Teece eds., 1992).

[31] *See, e.g.*, 1998 Biennial Regulatory Review—Spectrum Aggregation Limits For Wireless Telecommunications Carriers, *Notice of Proposed Rule Making*, 13 FCC Rcd. 25,132, ¶¶ 54-58 (1998).

[32] *Id.*

[33] Implementation of the Cable Television Consumer Protection & Competition Act of 1992, *Report & Order*, 14 FCC Rcd. 19,014 (1999).

current competition, but it also pledged to revisit the cap in two years. In the interim, it invited waiver requests from carriers that could show they were moving forward with new services that require additional spectrum. As already discussed, when the Commission did revisit the spectrum cap in 2001, it ordered the cap to be fully repealed by 2003 and to be raised to 55 MHz during the transition period.[34] In the cable ownership proceedings, the Commission imposed a subscriber limit.[35] But the FCC also said it would not attribute to an operator's subscriber count any customers to whom it provided only telephone or broadband services, (but not conventional cable video).

The effort in both of the cases above was to preserve competition without blunting incentives to invest in the development and deployment of new technology. The balance is an important one. If regulators or enforcement officials focus too rigidly on short-run competition and the immediate benefits of lower prices and higher output, they might in some cases place at risk longer-term benefits of innovation. The spectrum cap created precisely this kind of rigidity and its elimination brings the benefits of a more flexible, case-by-case approach to wireless mergers. But, if regulators too readily exchange actual competition for promised innovation, they risk creating market power without deriving any compensating benefit. For this reason, a rigorous antitrust approach to mergers in the mobile broadband markets is warranted.

Striking the right policy balance is especially challenging where, as with wireless telecommunications, technological change is a major and ongoing factor in the industry. The wireless market may be quite susceptible to what have been described as "waves" of innovation that transform not just individual firms, but an industry as a whole.[36] But, although maintaining or increasing existing market competition might have costs for innovation in specific cases, it is far less clear that such costs will often be at stake, even in the dynamic environment of mobile broadband services. Indeed, the available evidence suggests that competition policy for mobile broadband should hold a rebuttable presumption against claims that competition today must be sacrificed for deployment of innovative services tomorrow. The general empirical evidence on the relationship between market structure and innovation, and between firm size and innovation, is ambiguous. The data show no systematic relationship between the degree of market power of firms in

[34] 2000 Biennial Regulatory Review, *Report and Order*, 16 FCC Rcd. 22,668 (2001).

[35] Implementation of § 11 (c) of the Cable Act of 1992, *Report and Order*, 14 FCC Rcd. 19,098 (1999), *rev'd* Time Warner Entm't Co. v. FCC, 240 F.3d 1126 (D.C. Cir. 2001).

[36] JAMES UTTERBACK, MASTERING THE DYNAMICS OF INNOVATION: HOW COMPANIES CAN SEIZE OPPORTUNITIES IN THE FACE OF TECHNOLOGICAL CHANGE (1994).

an industry and the amount of innovative activity they undertake.[37] One study that focused specifically on the U.S. telecommunications industry, however, suggests a positive correlation between the speed with which firms deploy new technology in their networks and the amount of competition they face.[38] This evidence supports at least a starting presumption against allowing otherwise anticompetitive levels of consolidation in the name of innovation in the advanced wireless services market.

It is important to recognize that the case for careful merger scrutiny in dynamic markets does not translate into a case for breaking up, regulating, or penalizing monopolies that are honestly acquired and maintained in such markets. As the Supreme Court recently phrased a long-standing antitrust principle:

> The mere possession of monopoly power, and the concomitant charging of monopoly prices, is not only not unlawful; it is an important element of the free-market system. The opportunity to charge monopoly prices—at least for a short period—is what attracts "business acumen" in the first place; it induces risk taking that produces innovation and economic growth. To safeguard the incentive to innovate, the possession of monopoly power will not be found unlawful unless it is accompanied by an element of anticompetitive *conduct*.[39]

So long as such anticompetitive conduct does not occur, antitrust law counsels forbearance towards a firm that has worked its way to monopoly at the same time that it counsels scrutiny of two firms that try to merge their way to dominance.

D. Standard Setting in the Mobile Broadband Industry: Competing versus Common Platforms

The question of policy towards standardization in mobile broadband has several dimensions. Importantly, there is a global aspect to mobile broadband standard setting that can transcend the regulatory power of national competition policies. The European Telecommunications Standards Institute (ETSI) has made the adoption of a uniform wireless standard in Europe a principal policy goal. It was

[37] *See, e.g.*, Wesley M. Cohen & Richard C. Levin, *Empirical Studies of Innovation and Market Structure*, *in* HANDBOOK OF INDUSTRIAL ORGANIZATION (Richard Schmalensee & Robert Willig eds., 1989).

[38] Howard A. Shelanski, *Competition and Deployment of New Technology in U.S. Telecommunications*, 2000 U. CHI. LEGAL F. 85 (2000).

[39] Verizon Communications, Inc. v. Law Offices of Curtis V. Trinko, 124 S.Ct. 872, 879 (2004).

ETSI that adopted and then mandated implementation of Europe's second generation GSM standard.[40] ETSI has moved away from the underlying TDMA architecture of GSM for mobile broadband services, but has nonetheless backed a single W-CDMA mobile broadband standard known as the Universal Mobile Telecommunications Services (UMTS) standard.[41] The convergence to a single mobile broadband standard in Europe could have substantial consequences for mobile broadband standard setting elsewhere. For example, if the European market developed rapidly and a wide range of UMTS compatible handsets became available, then there might be incentives for mobile broadband providers in the United States or Asia to join the UMTS standard. To be sure, no such "tipping" towards a single network standard is necessary or inevitable, but under proper economic conditions, it is possible. The likelihood of tipping to a single standard increases if there are markets in which that standard is mandated, particularly if strong economic interests support regulatory perpetuation of the standard even as alternatives become available. Indeed, the prospect of anticompetitive results from a mandatory regional standard has been a central concern in the debate over standards policy for mobile broadband.[42]

At the global level, then, there is a competition policy question about the extent to which any governmental, or *de facto* governmental, body should mandate a standard. As things now stand, a variety of standards remain in global competition. The International Telecommunications Union (ITU) has accepted five standards that meet its "IMT-2000" criteria for roaming and data transport speed.[43] As a practical matter, three standards are viably competing in the mobile broadband market worldwide. The two major ones are UMTS, leading in Europe and Japan, and CDMA2000, which is strong in Korea and the United States. There is also a technology known as EDGE (Enhanced Data rates for Global Exchange), that will enable transition of TDMA and GSM-based 2G networks to mobile broadband capabilities.[44]

The fact that the mobile broadband standards race has boiled down to two or three options, and in some markets has converged to a single standard, does not signal the end of technological change in the wireless

[40] *See Standards in Wireless Telephone Networks*, *supra* note 11.

[41] *See id.*

[42] *See* Peter Grindley et al., *Standards Wars: The Uses of Standard Setting as a Means of Facilitating Cartels: Third Generation Wireless Telecommunications Standard Setting*, 3 INT'L J. COMM. L. & POL'Y 2 (1999).

[43] INTERNATIONAL TELECOMMUNICATION UNION, WHAT IS IMT-2000? (2001), *available at* www.itu.int/osg/imt-project/docs/What_is_IMT2000-2.pdf.

[44] *See* ITU Strategy and Policy Unit Newslog, *EDGE is a Competitive Tool* (Apr. 19, 2004), *at* www.itu.int/osg/spu/newslog/categories/mobile/2004/04/19.html.

market. The question going forward for competition policy is how standards should be set as wireless telecommunications evolve within 3G and beyond. There are three principal approaches: (1) government-coordinated standard setting, as with ETSI in Europe, (2) standard setting within private organizations, or (3) standards competition among individual firms.

The first option amounts to a blocking of standards outside those approved by the centralized body. This strategy might have short-run coordination benefits in the form of faster deployment and immediate compatibility but, as already mentioned above, is subject to a variety of hazards.[45] In particular, if the standards body is effectively controlled by particular interests such as powerful equipment manufacturers or the owners of particular intellectual property, then the centralized process could lead to entrenchment of a suboptimal standard that is, moreover, insulated from the competitive processes that could lead to its ultimate displacement through market forces. The policy choice then reduces to the question of whether or not to allow coordinated standard setting on a private basis by firms within the wireless industry.

In principle, there is no clear economic basis for an *ex ante* presumption for or against private standard-setting coalitions. Competition among standards spurs firms to innovate, to seek more effective and efficient technologies than their rivals have. Coalition around a sub-optimal standard may be less likely when standards are set competitively rather than cooperatively because multiple standards can be tested in the marketplace. Over time, prices decline and innovation may be encouraged under a competitive standards regime.

On the other hand, coalitions can lead to faster development of effective system standards and are more likely to achieve rapid compatibility among competing systems and complementary products. Commentators have attributed such virtues to the process that led to the GSM standard for "2G" wireless networks in Europe.[46] When system interfaces are standardized in an industry and are openly available to all firms at all levels within the industry, consumers can benefit from the resulting "mix and match" competition.[47] In addition, when standards are shared among competitors, price competition is likely to be intense as the rival firms will have more similar technologies and hence cost

[45] *See, e.g.*, Mark Lemley, *Standardizing Government Standard-Setting Policy for Electronic Commerce*, 14 BERKELEY TECH. L. J. 745 (1999).

[46] *See, e.g.*, Jacques Pelkmans, *The GSM Standard: Explaining a Success Story*, 8 J. EUR. PUB. POL'Y 432 (2001).

[47] *See* Jeffrey K. Mackie-Mason & Janet S. Netz, *Manipulating Interface Standards as an Anti-Competitive Strategy*, *in* STANDARDS AND PUBLIC POLICY (Victor Stango & Shane Greenstein eds., forthcoming 2005).

structures than may be the case under competitive standard setting. Standard-setting coalitions therefore have the potential benefit of inducing rapid diffusion of service and intense price competition. At the same time, however, they have the potential to impede competition by restricting membership, limiting access to the standard, and forcing industry adoption of the standard. This will be particularly true when the coalition includes firms with sufficient market power to impose a particular standard and excludes the most notable rivals to those firms.[48]

At a simplified level, one can cast the policy choice for standard setting as being between the short-run, static benefits of competition over a common standard and the dynamic innovation benefits of competition among rival system standards. That tradeoff makes the welfare effects of standards coalitions versus standards competition hard to predict. Indeed, American antitrust doctrine recognizes the potential benefits and ambiguous *ex ante* competitive effects of standard-setting organizations. It thus affords them "rule-of-reason" treatment rather than *per se* illegality under the Sherman Antitrust Act.[49] But the "static benefit versus dynamic benefit" characterization of the standards competition question is ultimately too simplistic. Importantly, standards-based competition does not necessarily result in competing standards. Competition among different standards may end in one technology's becoming dominant because of its objective superiority. Regulation and antitrust should not second-guess such outcomes. But a standard may also gain market power because of proprietary interfaces through which the owner can create feedback effects from complementary products and/or take advantage of network effects that deter users from switching to an alternative platform. Under certain conditions, such as where the network service provider also owns exclusive rights to the standard, the result could be the worst of all possibilities: a single standard but with only a single firm competing within that standard. This is unlikely in wireless communications where standards tend to be widely licensed by their developer(s). But where such a monopolistic outcome is possible a coalition might be preferable despite yielding only a single standard in the marketplace, because there would be several firms (the coalition members) competing within that standard.

At the same time, cooperative standard setting need not signal the end of innovation-based competition. There may be rival coalitions within the industry. New entrants may bring new standards into the market or some coalition members may defect to a superior standard.

[48] *See id.*

[49] *See* Allied Tube & Conduit Corp. v. Indian Head, 486 U.S. 492 (1988).

Indeed, even in the presence of strong network effects, economic analysis has shown that standards can change and networks can tip from one dominant technology to another.[50] In addition, a uniform standard at the platform layer of a network can spark increased innovation and competition in the applications layer. In the end, then, the welfare effects of standard-setting coalitions compared to standards competition are even more difficult to predict than the simple static-versus-dynamic-benefits story suggests.

So what, then, should competition policy be towards standard-setting coalitions among firms that otherwise compete in the relevant market? Antitrust authorities in this market should recognize (as they do in other contexts) that private standard-setting consortia can be beneficial and hence should not presume against their legality as a matter of competition law. Enforcement authorities should, however, be vigilant that coalitions do not structure themselves so as to gain power to act anti-competitively in their markets. The likelihood of anticompetitive outcomes from standards coalitions increases where membership is restricted and existing members determine who to admit to the coalition, where the coalition excludes important actual or potential competitors, and where the members of the coalition have sufficient market power to ensure industry adoption of their standard.[51] Competition policy should thus not be aimed at preventing the emergence of standards coalitions. But it should be applied to prevent standards consortia from operating as covers for group boycotts against certain competitors, or from serving as mechanisms by which owners of critical patents gain market power by forcing adoption of the standard to which their intellectual property rights are relevant.

E. Interconnection Among Competing Networks

An additional and related element of competition policy focuses not directly on standards, but on interconnection among rival networks. Even if competition policy does not take an initial position on how firms in the mobile broadband industry set system standards, law can have a profound effect on the competitive performance of the industry by requiring that subscribers to one system be able to trade traffic with subscribers on another, or by mandating that hardware devices used with mobile broadband be interoperable across competing technological platforms. Such interconnection policies have a notable history in the United States, sometimes more because of their absence rather than their

[50] Michael Katz & Carl Shapiro, *Systems Competition and Network Effects*, 8 J. ECON. PERSP. 93 (Spring 1994).

[51] *See* Mackie-Mason & Netz, *supra* note 47.

presence. It has become conventional wisdom, for example, that the absence of interconnection requirements in the early twentieth century allowed AT&T to squeeze out rival telephone companies and recapture the monopoly it had lost when its patents expired in the 1890's.[52] AT&T accomplished this by refusing to allow the rival network's customers to reach subscribers to AT&T's network. Because AT&T had the larger number of subscribers, its network was inherently more valuable to consumers because of the greater number of people one could call as a subscriber to AT&T than as a subscriber to any other network. This in turn attracted increasing numbers of customers to AT&T, which only increased and reinforced the strength of AT&T's advantage for consumers over other networks. The phenomenon whereby a service becomes more valuable to all users with each additional user of the service is often called a "network externality." The Telecommunications Act of 1996 instituted mandatory interconnection among competing carriers,[53] eliminating network externality advantages for incumbent carriers over new entrants.

The FCC extended interconnection to the wireless arena, requiring not only that wireless carriers interconnect with each other, but that wireline and wireless carriers also interconnect for the exchange of customer traffic.[54] The benefits that flow from mandatory interconnection are enormous and the lessons from existing wireless and wireline interconnection counsel that any competition policy towards mobile broadband services include such a mandate, a point on which there appears to be little debate. Such a requirement may, however, affect how standards are chosen and, if there are limits on interoperability among potential standards, tilt the process towards cooperative rather than competitive technological development. But as discussed above, so long as the cooperative standard setting is conducted in a non-exclusive manner and is not misused for the benefit of dominant firms, there is no reason for competition policy to stand in the way of standards coalitions. Similarly, if interconnection considerations lead service providers to converge on a standard owned by a single firm, the monopoly over the intellectual property rights to the standard should not give rise to concern so long as that monopoly is not maintained through anticompetitive strategies or misused to interfere with competition at the service level of the market.

[52] See BENJAMIN ET AL., *supra* note 26, at Ch. 15.

[53] *See* 47 U.S.C. § 251 (2000).

[54] Implementation of the Local Competition Provisions of the Telecomms. Act of 1996, *First Report & Order*, 11 FCC Rcd. 15,499 (1996).

F. Summary

In each of the four areas of competition policy discussed above, authorities must make difficult predictive judgments. In an evolving network industry like wireless telecommunications, factors affecting market definition, the feasible scope of competition, the relationship between market structure and innovation, and technological standards can all change rapidly. This section has attempted to anchor competition policy for mobile broadband services in fundamental antitrust principles that are responsive to the dynamic environment in which they are applied, but that retain a presumption in favor of preserving the most competitive market structure that is technologically and economically feasible. Therefore, the burden in each of the policy dimensions discussed should fall on parties seeking to engage in cooperative activity to prove that their conduct does not reduce competition or else has demonstrable efficiency or innovation benefits that offset the costs of reduced competition.

II. INSTITUTIONAL CONSIDERATIONS: WHO SHOULD IMPLEMENT COMPETITION POLICY FOR MOBILE BROADBAND?

Once the substantive framework for competition policy in the mobile broadband market is established, the institutional question arises of what kind of agency should implement that policy. Should competitive oversight lie with a general antitrust authority like the U.S. Department of Justice or FTC, or should it lie with a sector specific regulator like the FCC? In the United States, there has been a long history of shared authority between the FCC and the antitrust agencies over competition questions. For decades, the FCC had the greater level of authority and could even exempt mergers from scrutiny by the FTC or the Department of Justice.[55] The 1996 Act removed that exemption authority from the FCC and restored primary antitrust jurisdiction over telecommunications to the general antitrust agencies.[56]

The policy outlined above in this paper does not inexorably tend towards either a sector-specific telecommunications regulator or a general antitrust agency as the correct institution to oversee competition policy for mobile broadband, although it does favor implementation by the latter. Market definition, benchmarking, assessing innovation-based arguments, and examining standard-setting are exercises with which

[55] *See* Howard A. Shelanski, *From Sector-Specific Regulation to Antitrust Law for U.S. Telecommunications: The Prospects for Transition*, 26 TELECOMM. POL'Y 335 (2002).

[56] Communications Act of 1934, ch. 652, § 221(a), 48 Stat. 1048, 1080 (codifying the Willis-Graham Act, ch. 20, 42 Stat. 27 (1921)), *repealed by* Telecommunications Act of 1996, Pub. L. No. 104-104, § 601(b)(2), 110 Stat. 56, 143.

antitrust agencies are familiar and that they are well-equipped to handle. Indeed, each of the dimensions of competition policy discussed above is guideline-driven rather than rule-driven. There is no firm rule, like the spectrum cap, for determining the required market structure. There is instead the guideline that the market should not be allowed to concentrate to the point that firms achieve market power and cause long-run harm to consumers. Assigning competitive oversight to the Justice Department or the FTC would therefore be appropriate and in keeping with a U.S. trend towards moving competition policy for telecommunications away from the FCC and to the antitrust agencies.[57]

On the other hand, it is likely that some aspects of mobile broadband competition policy would be well governed by a sector-specific regulation. For example, the viability of competition among rival mobile broadband networks depends on interconnection among the networks for the purposes of exchanging calls among each other's subscribers. The oversight of interconnection and its associated pricing issues fits naturally with an agency like the FCC. Similarly, specific questions about standards or the usability of particular spectrum for entry into the mobile broadband market are also likely to be better addressed by an expert agency. Implementation of the policy framework outlined in this paper could therefore, in principal, afford a continued role to sector specific regulatory authorities. At the same time, however, this paper proposes an antitrust approach that should, for the most part, fall under the jurisdiction of general competition authorities.

CONCLUSION

This paper has examined the central dimensions of competition policy for mobile broadband services. The healthy development of 3G and even more advanced wireless capabilities is important in its own right. But it is also important because sound competition policy that promotes efficient development of the mobile broadband market will benefit consumers and help to mitigate the potential tradeoffs and market failures that underlie the regulatory debates over end-to-end neutrality for Internet transport networks, common versus licensed spectrum assignment, and open versus proprietary technological standards.

The premise of this analysis has been that competition policy should focus on protecting and enhancing consumer welfare in the relevant market. To that end, the principal dimensions of a competitive policy framework for mobile broadband should include (1) a conservative market definition that presumes inclusion only of mobile broadband

[57] *See id.*

mobile networks, but which cautiously takes account of potential substitutes and entrants in the uncertain and changing mobile broadband marketplace; (2) careful assessment of available spectrum and economies of scale to set an appropriate market-structure benchmark against which to assess competitiveness of the mobile broadband industry; (3) a wary approach to claims that dynamic innovation requires sacrifice of static competition, with the burden of persuasion resting with parties seeking market consolidation; (4) openness to private standard-setting coalitions coupled with vigilance for, and rigorous enforcement against, features of such organizations that might harm competition and accumulate market power; and (5) continued enforcement of interconnection for the exchange of traffic among competing networks.

Each of the above policy criteria lies squarely within the traditional ambit of antitrust law, suggesting that general antitrust agencies rather than sector-specific regulators should have the principal institutional role in applying competition policy to the mobile broadband industry. The above parameters of competition policy are broad and are susceptible to change given the nascent and dynamic nature of mobile broadband markets. But they constitute sound principles that, even if they must be applied flexibly over time, should provide a framework for fostering and preserving competition and consumer welfare in the evolving wireless marketplace.

EVOLVING CORE CAPABILITIES OF THE INTERNET

J. SCOTT MARCUS[*]

ABSTRACT

Historically, the Internet has served as an enormous hotbed of innovation. Nonetheless, deployment of a number of potentially beneficial and important Internet capabilities appears to be slowed or stalled for lack of sufficient commercial incentives. The primary concern is with *public goods*[1] where market forces alone might not be sufficient to drive widespread adoption. Timely and relevant examples are drawn primarily from the areas of network security and cybersecurity. How might government identify and prioritize those capabilities where intervention is warranted (if ever)? What actions on the part of industry and government are necessary and appropriate in order to ensure that societally significant problems, including network security and robustness, are addressed in the Internet?

[*] Author's current address: Federal Communications Commission (FCC), Office of Strategic Planning and Policy Analysis, 445 12th Street SW, Washington, DC 20554 and can be contacted at smarcus@fcc.gov. The author is affiliated with both the FCC and the European Commission, but the opinions expressed are solely those of the author, and do not necessarily reflect the views of either agency. The author is deeply indebted to his colleagues Richard Hovey and Jeffery Goldthorp, of the FCC; to Scott Bradner, of Harvard University; to Dale Hatfield, Gary Chapman and Andrew Johnson, of the University of Colorado; and to Scott Rose of the National Institute of Standards and Technology for a wealth of helpful and insightful comments.

[1] The Economist, *Economics A-Z*, ECONOMIST.COM, *available at* http://www.economist.com/research/Economics (last visited May 10, 2004) (adapted from MATTHEW BISHOP, ESSENTIAL ECONOMICS (2004).

TABLE OF CONTENTS

INTRODUCTION

Many have argued that the Internet is far more hospitable to innovation than the traditional public switched telephone network (PSTN).[2] Not so long ago, it seemed that all things were possible in the free-wheeling entrepreneurial and unregulated culture of the Internet. Nonetheless, it now appears that many seemingly promising innovations have languished in recent years. Is it possible that the Internet is hospitable to some innovations, but not to others? Is it possible that pure free market mechanisms will fall short in cases that are of vital importance to society at large? Might there be a role for government to play in promoting societally valuable goals that the market alone would not achieve? If so, what measures are available to government or industry to attempt to promote adoption of important and beneficial innovations?

One federal report, the draft version of *The National Strategy to Secure Cyberspace,* posed the key question succinctly: "How can government, industry, and academia address issues important and beneficial to owners and operators of cyberspace but for which no one group has adequate incentive to act?"[3] The final version of that same report offers an answer: "The government should play a role when private efforts break down due to a need for coordination or a lack of proper incentives."[4]

[2] *Cf.* David Isenberg, *The Rise of the Stupid Network*, COMPUTER TELEPHONY, Aug. 1997, at 16-26, *available at* http://www.hyperorg.com/misc/stupidnet.html.

[3] THE PRESIDENT'S CRITICAL INFRASTRUCTURE PROTECTION BOARD, THE NATIONAL STRATEGY TO SECURE CYBERSPACE*,* DRAFT FOR COMMENT 47 (2002), *available at http://www.iwar.org.uk/cip/resources/c-strategy-draft* [hereinafter DRAFT NATIONAL STRATEGY TO SECURE CYBERSPACE].

[4] THE PRESIDENT'S CRITICAL INFRASTRUCTURE PROTECTION BOARD, THE NATIONAL STRATEGY TO SECURE CYBERSPACE 31 (2003), *available at* http://www.whitehouse.gov/pcipb/cyberspace_strategy.pdf [hereinafter NATIONAL STRATEGY TO SECURE CYBERSPACE].

A particular concern here is with *public goods*. *The Economist* defines public goods as:

> Things that can be consumed by everybody in a society, or nobody at all. They have three characteristics. They are:
>
> - non-rival – one person consuming them does not stop another person consuming them;
> - non-excludable – if one person can consume them, it is impossible to stop another person consuming them;
> - non-rejectable – people cannot choose not to consume them even if they want to.
>
> Examples include clean air, a national defense system and the judiciary. The combination of non-rivalry and non-excludability means that it can be hard to get people to pay to consume them, so they might not be provided at all if left to market forces . . . . [5]

Most of the examples in this paper are drawn from the fields of network security and cybersecurity. In the aftermath of the events of September 11, 2001, there is a widespread recognition of the need to enhance the robustness and security of the Internet. Many security exposures exist. Techniques are available to prevent or at least mitigate the impact of the exploitation of certain of the known exposures; however, in certain instances, it is not clear that the organizations that would need to make investments to deploy the technologies are motivated to do so. This is especially likely where deployment costs would exceed the quantifiable economic benefits to the organizations that would have to bear those costs.

The Internet is unquestionably one of the greatest technological successes of modern times. Among the many factors that contributed to its success is the *end-to-end model*, which enables innovation at the edge of the network without changes to the core; and the absence of central control or regulation, which has enabled the Internet to evolve largely through private initiative, without the restrictions of cumbersome governmental oversight. To a large degree, the Internet represents a triumph of unbridled capitalist initiative.

Today, most networking professionals would agree that the Internet would benefit from a number of evolutionary changes – changes which, however, appear not to be forthcoming. In many cases, the technology

[5] The Economist, *supra* note 1. They go on to observe that, "public goods are regarded as an example of market failure, and in most countries they are provided at least in part by government and paid for through compulsory taxation." *Id.*

seems to be sufficiently straightforward, but deployment is stymied by a constellation of factors, including:

- the lack of sufficient economic drivers;
- the difficulty of achieving consensus among a plethora of stakeholders with interests that are either imperfectly aligned or else not aligned at all; and;
- the inability of government to foster change in an entity that is global in scope, and largely unregulated in most industrialized nations.

In other words, the very factors that fostered the rapid evolution of the Internet in the past may represent impediments to its further evolution. Historically, those Internet features that could be implemented through private initiative at the edge of the network emerged rapidly; those features that now require coordinated changes, and especially changes to the *core* of the network, are either slow to emerge or are not emerging at all.[6] One might now wonder whether the Internet has reached an evolutionary cul-de-sac.

This paper draws on examples associated with network security and cyber security; however, the issue of promoting public goods where market forces would otherwise be insufficient is a much larger topic. The author humbly asks the reader's indulgence as he frenetically jumps back and forth from the general to the more specific.

Readers who are well versed in the technology of the Internet may have an easier time following the issues, but this paper is not primarily about technology; rather, it focuses on the business, economic and regulatory factors that serve either to facilitate or to impede evolution. In any case, with the possible exception of Section II (which the reader could skip without loss of continuity), no prior knowledge beyond that of an intelligent layman is assumed as regards any of these disciplines.

This introduction provided a cursory overview of the issues. Section I provides background on factors that may militate against the deployment of certain kinds of enhancements to Internet functionality: the end-to-end principle, transaction costs, and the economics of network externalities (following the seminal work of Jeffrey Rohlfs).[7] Section II provides a brief technical overview of two emerging security

[6] *Cf.* Christian Sandvig, Communication Infrastructure and Innovation: The Internet as End-to-End Network that Isn't (Nov. 2002) (unpublished manuscript, available at http://www.cspo.org/nextgen/Sandvig.PDF).

[7] JEFFREY H. ROHLFS, BANDWAGON EFFECTS IN HIGH-TECHNOLOGY INDUSTRIES 3 (2001).

enhancements to the Domain Name Service (DNS), which collectively serve as an example of seemingly desirable security capabilities and the associated deployment challenges. Section III gingerly explores a topic that many in the Internet community will find uncomfortable: whether it is appropriate for government to play a more active role in fostering the further technical evolution of the Internet. Government intervention could be positive; it could be ineffective; or it could be counterproductive. What role, if any, should the U.S. Government play in the future technical evolution of the Internet? Section IV provides brief concluding observations.

I. BARRIERS TO ADOPTION

As part of the process of preparing the National Strategy to Secure Cyberspace, the President's Critical Infrastructure Protection Board (CIPB) convened a group of Internet experts. At a meeting of this group in May 2002, I commended them for their excellent and thoughtful recommendations.[8] I noted the importance of their work, and encouraged them to let their colleagues in government know if, as their work proceeded, they encountered difficulties in getting their firms to deploy the recommended facilities.

A moment of embarrassed silence followed. One of the attendees then timorously put up his hand and said:

> Scott, you don't have to wait a year or two to find out whether we are having problems getting this stuff deployed. We already know the answer. There is nothing new in these reports. All of this has been known for years. If we were able to craft business cases for our management, all of this would have been done long ago.

No one who has dealt with these issues in industry should be surprised by this answer. Certain Internet innovations have achieved widespread use with no market intervention, perhaps the most noteworthy being the World Wide Web. A great many other Internet innovations have languished, even though the underlying technology appeared to be sound.

[8] For a public summary of their major findings, see AVI FREEDMAN, AKAMAI TECHS., ISP WORKING GROUP INTERNET VULNERABILITY SUMMARY & DISCUSSION (2002), *available at* http://www.nanog.org/mtg-0206/avi.html.

In addition to the DNS security facilities described in this report, similar deployment concerns might be raised about:

- Internet Protocol (IP) version 6[9]
- Differentiated services (DiffServ)[10]
- IP multicast
- Operational tools and protocol enhancements to enhance the security of BGP-4 routing protocols

Engineers tend to conceptualize these deployment delays in terms of engineering concerns, such as incomplete protocol specifications, immature protocol software implementations, and insufficient interoperability testing. It may well be that these engineering problems are symptomatic of deeper business and economic impediments that militate against deployment and use of certain *kinds* of innovations in the Internet today.

This section of the paper discusses a constellation of economic factors that impede deployment of certain kinds of Internet facilities. The detailed interplay among these factors, and perhaps among other factors not considered here, may vary from one service to the next, but much of the observed behavior can apparently be explained by a small number of underlying economic factors.

A. Transaction Costs

Transaction costs are the economic costs associated with effecting a transaction.[11] Some transactions involve far higher transaction costs than others. If a customer buys a candy bar in a luncheonette, she typically hands the cashier some money, receives her change, and walks out the door with the desired item. Transaction costs are low. If that customer

[9] The National Telecommunications and Information Administration (NTIA), which is a part of the U.S. Department of Commerce, is currently conducting a Notice of Inquiry regarding IP version 6. Public comments are available at http://www.ntia.doc.gov/ntiahome/ntiageneral/ipv6/commentsindex.html. The parallels to DNS security are quite striking.

[10] Within the network of a single service provider, differentiated services are readily achievable. In the general, multiple-provider case, there is no significant deployment.

[11] Various definitions exist in the literature. *See, e.g.,* ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT, TRANSACTION COSTS AND MULTIFUNCTIONALITY, *available at* http://www1.oecd.org/agr/mf/doc/Transactioncosts32.pdf (last visited May 26, 2004) (citations omitted). It defines transaction costs in this way: "Transaction costs are 'the costs of arranging a contract *ex ante* and monitoring and enforcing it *ex post*' . . . 'the costs of running the economic system' . . . and 'the economic equivalent of friction in physical systems . . . .' " *Id.* at 2 (citations omitted).

purchases by credit card, the merchant pays a fee for the use of that credit card – transaction costs are higher. If a person buys or sells a house, transaction costs (broker's fees, loan initiation, and various fees) might consume a hefty 5-10% of the value of the transaction.

Transaction costs thus represent sand in the gears, a form of economic friction. Where a large number of parties must independently come to terms with one another on a single transaction, and particularly where those terms require substantial discussion or negotiation, transaction costs will tend to be very high.

High transaction costs cut into the *surplus* (the degree to which the value to a purchaser exceeds the cost) associated with a transaction. High transaction costs can literally be prohibitive – they can make the transaction as a whole uneconomic. Those who claim that the Internet is a hotbed of innovation are implicitly arguing that transaction costs to deploy new innovations on the Internet are low. In the pages that follow, this paper suggests that this is true only for certain kinds of innovations.

B. Network Externalities

The value of a network is largely a function of who can be reached over that network. Robert Metcalfe, the co-inventor of the Ethernet Local Area Network, attempted to roughly quantify this in *Metcalfe's Law*, which claims that the value of a network is roughly proportionate to the square of the number of users.[12]

Most industries experience economies of scale – bigger is better. Networks, however, are subject to additional effects of scale that go far beyond traditional economies of scale. Every time that someone in North Dakota obtains telephone service for the first time, it enhances the value of *everyone's* telephone service – there is one more person who can be reached by phone. Economists refer to these effects as *network externalities*, or informally as *bandwagon effects*.

For a product or service subject to substantial network externalities, nothing succeeds like success. One of the most common examples of a bandwagon effect is the competitive clash of two videocassette standards, VHS and Betamax. At a technical level, neither had a decisive advantage over the other, and for a time they coexisted in the marketplace. Over time, VHS acquired more customers. As a result, studios developed more programming in the VHS format. Consumers with Betamax

[12] *Cf.* Andrew Odlyzko, *Content is Not King*, FIRST MONDAY, Jan 8, 2001, *at* http://www.firstmonday.dk/issues/issue6_2/odlyzko/ (arguing that ". . .Metcalfe's Law does not reflect properly several other important factors that go into determining the value of a network. However, the general thrust of the argument . . . [is] valid.").

equipment found less and less of interest in rental stores, and eventually nothing at all. "Eventually, all consumers – even those who preferred Beta[max]'s picture quality . . . – had no choice but to get on the VHS bandwagon."[13]

In some instances, network externalities manifest themselves by way of direct interactions with other users of the same network. In others, the bandwagon effects relate to complementary upstream or downstream industries, as was the case with VHS and Betamax (the player was valuable only if extensive content was available to play on it). These complementarities often lead to the classic "chicken and egg" problem, where two vertically related industries cannot succeed unless both are launched at once.

In a bandwagon marketplace, multiple stable equilibria are usually possible, and these equilibria can differ greatly. Rohlfs defines the *initial user set* as comprising "all individual entities . . . that can justify purchasing the service, even if no others purchase it."[14] If the demand for the service is enhanced by being taken up by the initial user set, then additional users will acquire the service until a higher equilibrium is reached, the *demand-based equilibrium user set*. The level of usage that is societally optimal, the *maximum equilibrium set*, may be much larger than the demand-based equilibrium user set.[15]

Unfortunately, "ordinary demand adjustments do not provide a path to the optimum."[16] Achieving the maximum equilibrium set often requires "supply-side activities or government intervention."[17]

New technology products and services have to get over an initial "hump" in order to reach critical mass. Different high-technology industries have achieved critical mass in different ways. Large numbers of videocassette recorders (VCRs) were sold to time-shift television programs on a stand-alone basis; subsequently, these VCRs established the necessary preconditions for the videocassette rental business that today represents the primary use of the VCR.[18] For CD players, necessary complementary products became available due to vertical integration – the same firms that were manufacturing CD players (Phillips and Sony) had significant ownership interests in producers of recorded music.[19] For black and white television, industry convergence on the National Television Standards Committee (NTSC) technical

[13] ROHLFS, *supra note* 7. (The discussion of network externalities that follows draws heavily on Rohlfs's work.).

[14] *Id.* at 23.

[15] *Id.* at 24.

[16] *Id.*

[17] *Id.*

[18] *Id.* at Ch. 10.

[19] ROHLFS, *supra note* 7, at Ch. 9.

standard, coupled with its rapid adoption by the FCC, played a large role in overcoming the initial start-up problem.[20]

C. Misalignment of Incentives

In a largely unregulated, market-based system, firms make business decisions based on anticipated costs and benefits. Any decision to change a firm's existing operating environment will entail initial costs. If the firm is to incur those costs, it must believe that there will be corresponding benefits that exceed those costs.

A recent report by the Institute for Infrastructure Protection (I3P) describes the dilemma:

> In a market-based economic system, it is not surprising that the market for IT and cyber security products defines the state of cyber security. Two closely related questions appear to drive decisions on how security products and services are acquired and used: (1) what are the cyber security risks to the enterprise and how do they fit into the overall risk equation of a company, and (2) what is the value of cyber security – how much financial benefit it provides. There are no clear answers to these questions.[21]

Features that constitute public goods (such as enhancements to network security) do not in general reduce recurring operating costs, so the benefits must come from somewhere else. Many organizations find it difficult to justify these expenditures for one or more of a number of reasons. Notably, the benefits may be difficult or impossible to quantify,[22] or whatever benefits exist may accrue to a party or parties other than the firm that must make the investments. Collectively, these two factors mean that the organization is unlikely to be motivated to make the investment.

[20] *Id.* at Ch. 12.

[21] INSTITUTE FOR INFORMATION INFRASTRUCTURE PROTECTION, CYBER SECURITY RESEARCH AND DEVELOPMENT AGENDA 40 (2003), *available at* http://www.thei3p.org/documents/2003_Cyber_Security_RD_Agenda.pdf [hereinafter I3P REPORT].

[22] *Id.* at 34-45.

> Decision makers lack a foundation of data about the current investment and risk levels: metrics that express the costs, benefits, and impacts of security controls from an economic perspective, technical perspective, and risk perspective; and ways to predict the consequences of risk management choices. . . . Risk assessment and dependency modeling for cyber security remain in an immature state with only little momentum in the marketplace.

Id.

D. The Time Frame of Risks and Rewards

Après moi, le déluge! (After me, the flood!)[23]

Firms fund business cases where the expected return exceeds the expected investment within some defined period of time.

Many cyber vulnerabilities relate to potential exploits that have very high cost, but very low probability of occurrence. These are "thirty year flood" events. Firms may resist funding solutions to thirty year flood problems for some combination of reasons, including:

- The business case takes too many years to prove in;
- The risks are too speculative, and thus too difficult to quantify;
- The risks are born primarily by their insurers, or possibly by the government;
- They may believe, rightly or wrongly, that even if the event takes place, they are unlikely to be viewed as negligent if their competitors were similarly unprepared;
- The current managers may consider it unlikely that the event will happen while they are still with the firm. They bequeath the problem, if indeed it proves to be a problem, to their successors.

E. The TCP/IP Reference Model

The underlying architecture of the Internet has significant implications for the transaction costs associated with the deployment of new capabilities. This part of the paper describes the architecture of the Internet in order to motivate the discussion of the economics associated with the *end-to-end principle* that appears in the subsequent section.

Perhaps the most significant advance of the past thirty years or so in data networking is the advent of *layered* network architectures. A layered network architecture breaks the functions of a data network up into functional layers, each of which communicates with its peer layers in other communicating systems, while deriving services from the layer

[23] Attributed to Louis XV, king of France from 1715-1774. Some sources instead attribute this quotation to his mistress, Madame de Pompadour.

beneath. This layering helps insulate one layer from another, providing many benefits – a topic we return to later in this section of the paper.

The TCP/IP protocol family, or *protocol suite*, is the preeminent example today of such a layered network architecture.[24] The TCP/IP protocol suite is based on a conceptual model that characterizes the communications hardware and software implemented within a single communicating system – for instance, the personal computer (PC) on your desk – as being comprised of a protocol stack containing multiple layers (see Figure 1).[25]

Levels 1 and 2, the *Physical* and *Data Link Layers* respectively, represent the realization of the "wire" over which communication takes place and the management of that wire. For instance, the Data Link Layer might determine which of several computers is authorized to transmit data over a particular local area network (LAN) at a particular instant in time.

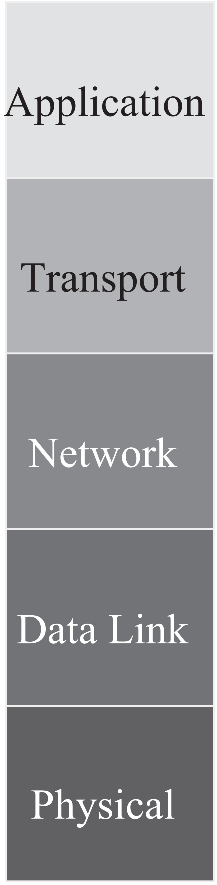
Level 3, the *Network Layer*, forwards data from one interconnected network to the next. For the Internet, the Network Layer is the *Internet Protocol* (IP)*,* which independently routes and forwards small units of data (datagrams).

Level 4, the *Transport Layer*, processes those datagrams and provides them to whichever application needs them, in the form that the application requires. For the Internet, the *Transmission Control Protocol* (TCP) supports applications that need a clean and reliable stream of data with no omissions or duplicates. The *User Datagram Protocol* (UDP) represents an alternative Transport Layer protocol that supports applications that do not require the tidy delivery that TCP provides. E-mail uses TCP, while Voice over IP (VoIP) uses UDP.

[24] The evolution of the TCP/IP protocol suite was influenced by earlier layered network architectures, and influenced in turn the subsequent evolution of a number of those network architectures. Among the layered network protocol families that emerged during the Seventies and Eighties were CCITT's X.25, IBM's System Network Architecture (SNA), Digital Equipment Corporation's DECnet, and Xerox Network Systems (XNS). Perhaps the most influential layered network architecture was the Reference Model for Open Systems Interconnection, usually referred to as the *OSI Reference Model*. The OSI Reference Model was developed jointly by the International Organization for Standardization (ISO) and the ITU/CCITT. The most readable descriptions of the OSI Reference Model appear in Hubert Zimmerman, *OSI Reference Model – The ISO Model of Architecture for Open Systems Interconnection*, 4 IEEE TRANSACTIONS ON COMM. 425 (1980), and in ANDREW TANENBAUM, COMPUTER NETWORKS (Prentice Hall 3d ed. 1996).

[25] Rigid adherence to protocol layering tends to impose a high overhead on protocol software. In reality, TCP/IP implementations often combine layers or take short-cuts as a means of reducing this overhead. *See* DAVID D. CLARK, RFC 0817: MODULARITY AND EFFICIENCY IN PROTOCOL IMPLEMENTATION (Internet Engineering Task Force, July 1982), *at* http://www.ietf.org/rfc.html.

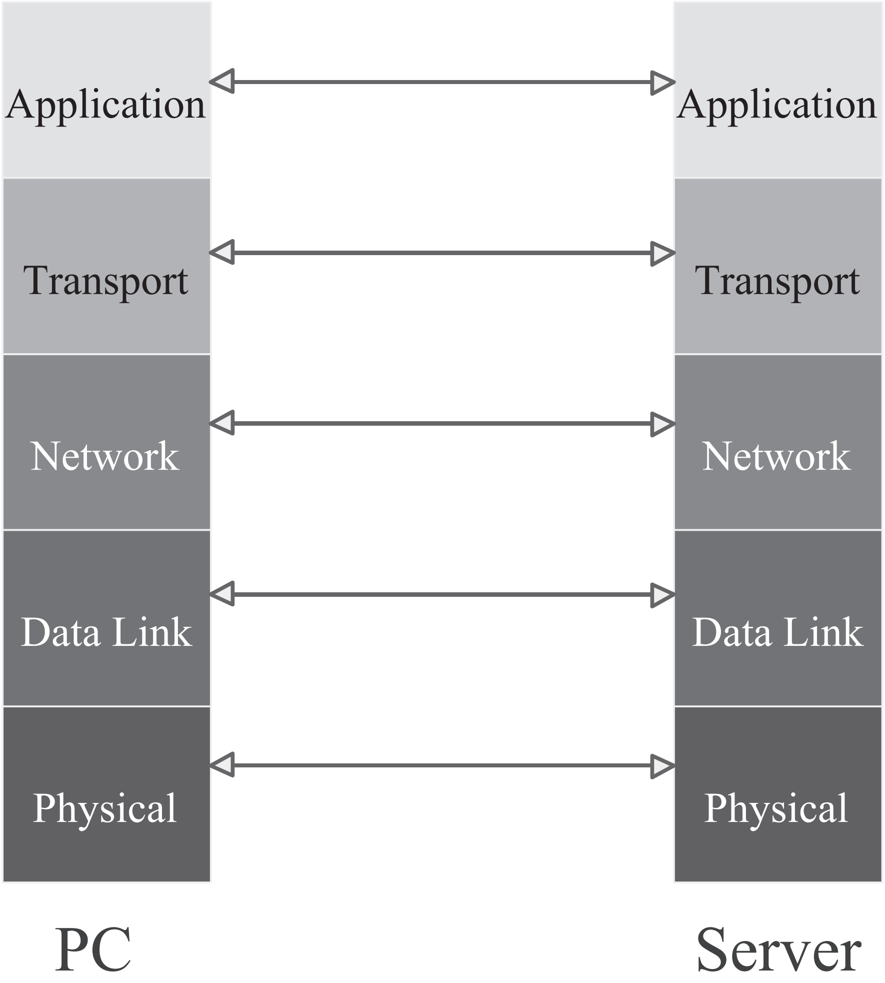
FIGURE 1
PROTOCOL LAYERS IN THE OSI / INTERNET REFERENCE MODEL



Level 5, the *Application Layer*, performs useful work visible to the end user, such as the browser or e-mail client (SMTP, HTTP) on your PC.

In this reference model, a layer logically interacts with its peer in a communicating system (see Figure 2). Thus, an Application Layer, such as the web browser in your PC, communicates with its peer process, a web server in a distant computer.

FIGURE 2
PEER LAYERS LOGICALLY INTERACT WITH ONE ANOTHER



Each layer within a communicating system implements this logical interaction by requesting services from the next lower layer. Thus, the Application Layer requests data from the Transport Layer. In doing so, it uses an interface that intentionally hides the details of how the lower layer implements its service. This information hiding is a key beneficial property of a layered network architecture – it enables the implementation of a layer to change without impacting the layers above or below.

FIGURE 3
LOGICAL AND PHYSICAL INTERACTIONS BETWEEN NETWORK PROTOCOL LAYERS

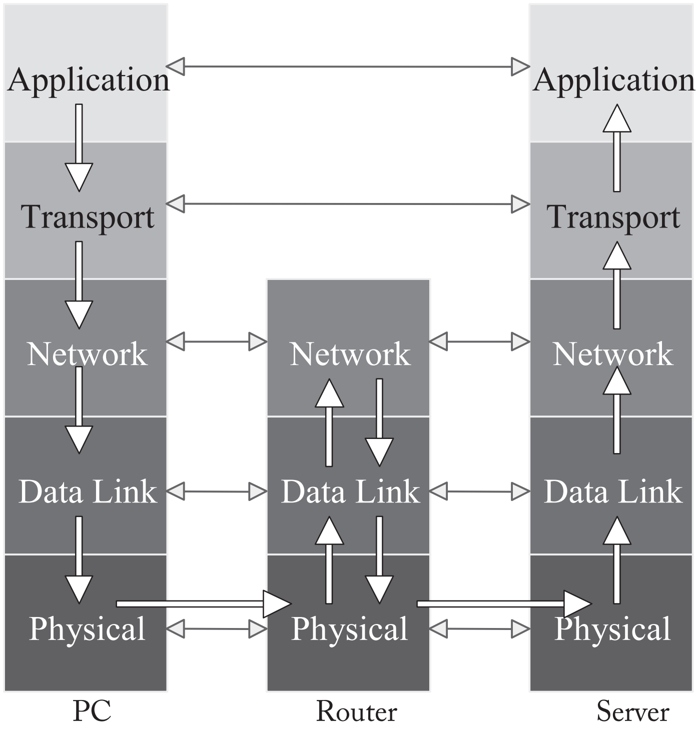


Figure 3 shows the relationship between logical and physical interactions in the Internet layered network architecture. It also adds another element to our understanding – a *router*, which is a device that exists solely to forward traffic in the Internet.

The information hiding property of a layered network architecture facilitates technical innovation over time. It also enables network applications to be written once to operate over any underlying transmission technology, or combination of technologies, thus simplifying the application creator's job. Conversely, the creator of a new transmission technology need only ensure that adequate interfaces exist to enable upper layers of the network to *use* the new communications layer – there is no need to make network applications specifically *aware* of a new underlying transmission technology. Phrased differently, a new network application will work with existing networks, and no changes are needed to underlying network transmission technologies. A new network transmission technology will work with existing networks, and no changes will be needed to existing applications. These properties greatly simplify the evolution of the network over time, and thereby reduce the transaction costs associated with network evolution.

F. The End-to-End Principle

In the early Eighties, a number of distinguished computer scientists at MIT propounded the *end-to-end principle*.[26] They noted that certain communications capabilities were most appropriately associated, not with the underlying network, but rather with the application that used the network. End-to-end reliability of transmission, for instance, could truly be assured only at the end points themselves. They further argued that, if the function could only be correctly implemented in the end points of the network, that it was a bad idea to also implement these functions in intermediate systems—doing so introduced not only inefficiencies, but also an increased possibility of error. Internet engineers have generally accepted the end-to-end principle as a basic tenet of network design. Moreover, they have sometimes advanced the further argument that the end-to-end principle fosters the evolution of the Internet, in that it enables new applications to be developed at the edges of the network, without disrupting the underlying core.[27]

There is much to be said for this view. For example, the creation of the World Wide Web initially depended primarily on the creation of a browser that could read and interpret existing file formats, and secondarily on servers for HTTP. No prerequisite changes were needed to the underlying TCP/IP protocols, the IP addressing system, or the DNS—these already provided the necessary support. This absence of prerequisite changes in turn reduced the number of parties that had to change their infrastructure – no action was required, for instance, on the part of Internet Service Providers (ISPs). By reducing the number of parties who must act in order to implement a particular change to the Internet, the end-to-end principle reduces the transaction costs associated with the development of new applications, thus fostering the continuing evolution of the Internet.[28]

[26] J.H. Saltzer et al., *End-to-End Arguments in System Design*, *in* ACM TRANSACTIONS ON COMPUTER SYSTEMS 2, 277 (1984), *available at* http://web.mit.edu/Saltzer/www/publications/endtoend/endtoend.pdf.

[27] Isenberg, *supra* note 2.

[28] For an interesting economic interpretation of the costs and benefits of this flexibility, see Mark Gaynor et al., *The Real Options Approach to Standards for Building Network-based Services* (2nd IEEE Conference on Standardization and Innovation in Information

More recently, a revisionist scholar, Christian Sandvig, has called this view into question.[29] He notes that this interpretation of the end-to-end principle presupposes that the underlying network already provides all of the functionality that will ever be necessary or desirable. In fact, it is difficult to know the impact of "missing" functionality – people develop applications to fit the functionality that is already available. Nobody takes the time to develop the applications that would have failed due to insufficient support in the underlying network; consequently, there is no obvious "graveyard" of failed applications.

Thus, while the end-to-end principle may tend to facilitate the development of new data networking *applications* (based in the Transport thru Application Layers of the familiar OSI Reference Model,[30] as described earlier in this paper),[31] it does nothing to foster the evolution of the underlying functionality associated with the Network Layer and below.

As it happens, this same OSI Reference Model has largely succeeded in decoupling and simplifying the evolution of its lowest layers. Below the Network Layer – which for TCP/IP is the Internet Protocol – datagrams can be transmitted over any Data Link Layer that is known to two systems that are topologically[32] adjacent. This is so because the lowest layers, the Physical and Data Link Layers, operate on a *point-to-point* basis.

Some years ago, the Dutch logician Edsgar Dijkstra conceived the notion of *structured programming*.[33] By a clean nesting of logical functionality, it was possible to contain the impact of changes to a program to a defined scope of statements within the program. This greatly enhanced the reliability of programs, and made it much easier to evolve programs (because a change in one part of the program was unlikely to cause unexpected and unpredictable adverse impact somewhere else).

A similar evolution took place for database management systems – by segregating functionality into various *schemas*, and hiding unnecessary details about how those schemas implemented their

Technology, Oct. 2001), *available at* http://people.bu.edu/mgaynor/papers/IEEE-standard-camera.pdf.

[29] Sandvig, *supra* note 6.

[30] Zimmerman, *supra* note 24 (the TCP/IP protocol suite that forms the foundation of the Internet broadly follows the OSI Reference Model, but with simplification in the upper layers).

[31] *See supra* Section I.E.

[32] Topology is the branch of mathematics that deals with the interconnectivity of the vertices and edges that comprise geometric figures, without considering their dimensions. It provides a useful way to visualize communications networks and to express their formal properties.

[33] O.J. DAHL ET AL., STRUCTURED PROGRAMMING (1972).

respective functions, the database systems fostered greater reliability and ongoing functional evolution.

The OSI Reference Model attempted to apply similar principles to data networks. The functionality of the network was broken down into seven functional layers (five for the TCP/IP world). The upper layers were associated with the application, the lower layers with the transmission mechanism. Each layer communicated with its peer layer in another communicating system; however, each effectuated this communication by requesting services from the layer beneath it. A layer never needed to know *how* the underlying layer provided the functionality.

There is no need for the entire Internet to understand any particular Data Link protocol mechanism. A given system that participates in the Internet need only understand those Data Link protocols whereby it communicates with the systems with which it maintains direct point-to-point communications. These systems could be said to be *topologically adjacent*.

These properties provide a decoupling for the lower layers of the OSI Reference Model that is very similar in effect to that which the end-to-end principle provides for the upper layers. New applications can be implemented as communicating processes in any two cooperating systems. Likewise, new transmission facilities at the Data Link Layer and below can be implemented in any two adjacent cooperating systems. In both cases, the transaction costs associated with deployment are bounded.

All of this breaks down for the Network Layer, IP. IP provides global connectivity and interoperability for the Internet. There are, of course, ways to evolve the IP functionality of the Internet, but these tend to be complex. There is no assurance that a change made between a pair of systems will have no impact on other systems. There is no inherent mechanism for information hiding within the IP Layer. Any functional evolution must be orchestrated with exquisite caution, because there is no guarantee that the unintended consequences of a given change will be limited.

In sum, technology evolution tends to be complex and expensive for the IP Layer, and also for certain other elements of the Internet that are global in scope. Since the transaction costs associated with evolutionary change of these elements are high, the benefits of any proposed evolutionary change would have to be correspondingly high – otherwise, the deployment of the proposed change is likely to stall for lack of a sufficiently compelling business case.

II. THE TECHNOLOGY OF DNS SECURITY

There are a wide variety of Internet facilities that might logically fall within the scope of this discussion. In order to motivate the discussion, we focus on a specific constellation of potential Internet security features associated with the DNS.

This paper does not attempt to argue whether any particular Internet security service is in some sense essential. Rather, the intent is to provide background on the rationale of a particular Internet service whose relatively slow deployment might in some sense be emblematic of a broader issue, to assume *arguendo* that there were some pressing requirement for deployment of that service, and then to pose the question: What impediments to deployment are visible today, and what further impediments might we anticipate in the future? By conducting this thought exercise, we come to a better understanding of the challenges that any feature of this type is likely to encounter.

In this sense, DNS security serves merely as a plausible proxy for any of the Internet-based services that we might have considered.

A. The Domain Name System

The DNS is the primary mechanism whereby names, such as www.fcc.gov, are mapped to Internet addresses, such as 192.104.54.3. The DNS has other mapping or directory functions as well.[34]

A DNS *client,* which might reside in your PC, initiates a DNS request to determine the IP address of www.fcc.gov. The request might be sent to a DNS server maintained by a company or by an ISP, the firm that provides access to the Internet.

The DNS is usually thought of as representing a logical tree structure. The root of that tree is comprised of thirteen groups of DNS servers in the United States, Europe and Asia.[35] Below the root are other groups of servers associated with Top Level Domains (TLDs), which are

[34] The DNS is documented in a series of Requests for Comments (RFC) that were developed by the Internet Engineering Task Force (IETF). The primary references are P.V. MOCKAPETRIS, RFC 1034: DOMAIN NAMES - CONCEPTS AND FACILITIES (Internet Engineering Task Force, Nov. 1, 1987), *at* http://www.ietf.org/rfc.html [hereinafter *RFC 1034*] (updated by *RFC 1101*, *RFC 1183*, *RFC 1348*, *RFC 1876*, *RFC 1982*, *RFC 2065*, *RFC 2181*, *RFC 2308*, *RFC 2535*); and P.V. MOCKAPETRIS, RFC 1035: DOMAIN NAMES - IMPLEMENTATION AND SPECIFICATION (Internet Engineering Task Force, Nov. 1, 1987), *at* http://www.ietf.org/rfc.html [hereinafter *RFC 1035*] (updated by *RFC 1101*, *RFC 1183*, *RFC 1348*, *RFC 1876*, *RFC 1982*, *RFC 1995*, *RFC 1996*, *RFC 2065*, *RFC 2136*, *RFC 2181*, *RFC 2137*, *RFC 2308*, *RFC 2535*, *RFC 2845*, *RFC 3425*, *RFC 3658*). All RFCs are available at http://www.ietf.org/rfc.html.

[35] Some of these root servers are now mirrored in multiple locations.

associated with the rightmost portion of a domain name[36] – for example, .com, .org, or .gov. The servers responsible for .gov provide in turn pointers to the next level down, including servers responsible for .fcc.gov.

This tree structure facilitates delegation of authority.

B. Security Exposures in the DNS

> The opening word was inscribed on the archway all the time! The translation should have been: *Say 'Friend' and enter*. I had only to speak the Elvish word for *friend* and the doors opened. Quite simple. Too simple for a learned loremaster in these suspicious days. Those were happier times.[37]

The DNS was designed in happier times, with little or no regard for security concerns.[38] When a DNS request is transmitted, there is no assurance that the response came from the desired DNS server, nor that the information provided was valid.

If a malefactor (who somehow had the ability to eavesdrop on DNS requests for the address of www.fcc.gov) wished to subvert the FCC's web site, they would not need to hack www.fcc.gov; they could instead create their own bogus site, and respond to DNS requests with the IP address of the bogus site. They might not even have to block legitimate DNS responses; it would be sufficient to respond faster than the legitimate DNS servers. Users accessing the bogus site would presume it to be the real one. There are countless variants on this scenario. Most of them depend on one of several underlying exposures:[39]

[36] Strictly speaking, we should say the rightmost *customarily visible* portion of the domain name. The rightmost portion is a period denoting the root itself, which is unnamed; however, this is often omitted by convention.

[37] J.R.R. TOLKIEN, THE FELLOWSHIP OF THE RING 402 (Ballantine Books 1965).

[38] *Cf.* I3P REPORT, *supra* note 21, at iii ("The information infrastructure, taken as a whole, is not an engineered system. . . . Security was not a significant consideration at its inception, and security concerns today do not override market pressures for new uses of technology or innovation, in spite of frequent stories of hackers, criminals, and, increasingly, terrorists and nations using or planning to use the information infrastructure as a weapon to harm the United States.").

[39] *Cf.* D. ATKINS & R. AUSTEIN, RFC __: THREAT ANALYSIS OF THE DOMAIN NAME SYSTEM (Internet Engineering Task Force, Feb. 2004), *at* http://www.ietf.org/internet-drafts/draft-ietf-dnsext-dns-threats-07.txt (work in progress: RFC is in preparation). Atkins and Austein primarily characterize threats as (1) packet interception, (2) ID guessing and query prediction, (3) name games, (4) betrayal by trusted server, and (5) denial of service. *Id.* Much work has been done over the years to characterize threats to the DNS, notably including Steven Bellovin, *Using the Domain Name System for System Break-Ins*, USENIX, (Jun. 1995), *at* http://www.usenix.org/publications/library/proceedings/security95/bellovin.html.

- There is no *authentication* of the DNS server, i.e. no assurance that the server is who it purports to be;

- There is no assured *integrity* of the DNS response, i.e. no assurance that the message received is the same as that which was sent;

- There is no assurance that the data maintained by the DNS server was not somehow maliciously modified on the server before being sent. There is in any event no assurance that the data is correct;

- Because the DNS is a logical tree, any compromise potentially impacts everything below that point in the DNS tree.

There is also concern that malefactors might attempt to cripple large portions of the Internet by launching *Distributed Denial of Service* (DDoS) attacks against key DNS servers, preventing users from reaching DNS servers. If users cannot resolve certain domain names, then to all intents and purposes they are unable to use the Internet to access those computers. An attack that was launched on October 21, 2002 received considerable media attention. All indications are that the October 21 attacks had minimal impact; nonetheless, the attacks demonstrated that denial of service is a real threat whose impact should not be underestimated.

C. DNS Security Mechanisms

The Internet community has been aware of these security exposures for many years. A number of responses have been developed within the *Internet Engineering Task Force* (IETF), the relevant standards body. Some of these are potentially more effective than others.

An exhaustive description of these systems is beyond the scope of this paper. The reader who desires more detail should consult the relevant Internet Request for Comments (RFC) documents. I provide a very brief summary here.

1. Domain Name System Security Extensions

The primary response to these security exposures has been the development of a series of specifications for Domain Name Security Extensions,[40] notably *RFC 2535*, that are sometimes termed *DNS Security Extensions* (DNSSEC).[41]

[40] DONALD EASTLAKE III, RFC 2535: DOMAIN NAME SYSTEM SECURITY

RFC 2535 provides for the storage of public cryptographic keys as a new DNS resource record. Keys are used both to authenticate the data's origin, and to assure the integrity of an RRset (a set of DNS resource records).

The authentication mechanism depends on the establishment of a *chain of trust*. The chain flows from the root of the DNS system (or from some other point in the DNS tree that is by convention assumed to be trustworthy) down to individual DNS leaf entries. The intent is that DNS servers would intrinsically and reliably be aware of the key for the root zone, and would follow trusted and authenticated entries through each level of the DNS tree in order to reach the correct leaf.[42]

The creators of *RFC 2535* were also concerned about the possible exploitation of negative information in the DNS – responses erroneously claiming that a domain name does *not* exist. Given that the domain name space is sparse, merely signing the entries that are present would not necessarily prove that a domain name did not exist. *RFC 2535* as amended addresses this by providing for an NSEC resource record[43] which points to the next valid domain name in what we can loosely term alphabetical order.

RFC 2535 is currently an IETF Proposed Standard. This means that it "is generally stable, has resolved known design choices, is believed to be well-understood, has received significant community review, and appears to enjoy enough community interest to be considered valuable."[44]

EXTENSIONS (Internet Engineering Task Force, Mar. 1999), *at* http://www.ietf.org/rfc.html (updated by *RFC 2931*, *RFC 3007*, *RFC 3008*, *RFC 3090*, *RFC 3226*, *RFC 3445*, *RFC 3597*, *RFC 3655*, *RFC 3658*) [hereinafter *RFC 2535*]; DONALD EASTLAKE III, RFC 2541: DNS SECURITY OPERATIONAL CONSIDERATIONS (Internet Engineering Task Force, Mar. 1999), *at* http://www.ietf.org/rfc.html [hereinafter *RFC 2541*].

[41] To avoid confusion, we use the term "*RFC 2535* DNSSEC" to refer specifically to *RFC 2535* capabilities. Some sources use DNSSEC to refer only to *RFC 2535*, while others use it to encompass additional capabilities, including TSIG, secure dynamic updates (per *RFC 3007*), and the CERT resource record (*RFC 2538*).

[42] This seemingly simple assumption masks a world of complexity. For example, the root signature, like all signatures, should be periodically changed in case it has been somehow compromised, and also to minimize the risk of cryptanalysis. If the key is statically configured in every client, how can it reliably be updated? *See RFC 2541, supra* note 40. *See also RFC 2535*, *supra* note 40, at § 6.2.

[43] In the original *RFC 2535*, the corresponding RR was referred to an NXT resource record. Based on operational experience, a number of non-backward-compatible changes were made to the DNSSEC protocols, culminating in a renaming of several RRs and renumbering of their code points. See S. WEILER, RFC 3755: LEGACY RESOLVER COMPATIBILITY FOR DELEGATION SIGNER (DS) (Internet Engineering Task Force, May 2004), *at* http://www.ietf.org/rfc.html [hereinafter *RFC 3755*].

[44] SCOTT BRADNER, RFC 2026: THE INTERNET STANDARDS PROCESS –REVISION 3, § 4.1.1 (Internet Engineering Task Force, Oct. 1996), *at* http://www.ietf.org/rfc.html [hereinafter *RFC 2026*].

At the same time, early operational tests have raised questions about a number of important protocol details.[45]

RFC 2535 provides for a very comprehensive any-to-any security mechanism, but it is operationally and computationally relatively expensive. There is a natural tendency to focus solely on the incremental cost of hardware and software, but the relevant deployment costs also include training; deployment planning, testing and staging; and ongoing operational complexity and associated incremental expense. Initial generation of public/private key pairs is computationally intensive, as is periodic or episodic re-signing of a DNS zone. Validation of signatures by means of public key cryptography is also computationally intensive – far more so than private key cryptography. The use of *RFC 2535* increases the length of DNS responses, and greatly increases the size of the DNS database.[46] Ultimately, the cost of increased computational power and server storage may be less important than the incremental expense associated with a substantial increase in operational complexity – ensuring the secrecy of the private keys, and effecting re-signing without breaking the chain of trust are just a few examples.[47]

2. Secret Key Transaction Authentication for DNS (TSIG)

A second response has been the use of TSIG to validate, for example, zone transfers[48] (the transfer *en masse* of a possibly large

[45] For more information on this topic, visit RIPE NCC, DEPLOYMENT OF INTERNET SECURITY INFRASTRUCTURES, *at* http://www.ripe.net/disi/ (last visited May 26, 2004).

[46] One source claims that it increases the size of the DNS database by a factor of seven. *See* PAUL ALBITZ & CRICKET LIU, DNS AND BIND 308-74 (4th ed. 2001), *available at* http://www.oreilly.com/catalog/dns4/chapter/ch11.html.

[47] *Id.* at 374 ("We realize that DNSSEC is a bit, er, daunting. (We nearly fainted the first time we saw it.)").

[48] P. MOCKAPETRIS, RFC 1034: DOMAIN NAMES – CONCEPTS AND FACILITIES § 4.3.5 (Internet Engineering Task Force, Nov. 1987), *at* http://www.ietf.org/rfc.html [hereinafter *RFC 1034*]. *RFC 1034*, describes DNS zone transfers in this way:

> "Part of the job of a zone administrator is to maintain the zones at all of the name servers which are authoritative for the zone. When the inevitable changes are made, they must be distributed to all of the name servers. While this distribution can be accomplished using FTP or some other ad hoc procedure, the preferred method is the zone transfer part of the DNS protocol. The general model of automatic zone transfer or refreshing is that one of the name servers is the master or primary for the zone. Changes are coordinated at the primary, typically by editing a master file for the zone. After editing, the administrator signals the master server to load the new zone. The other non-master or secondary servers for the zone periodically check for changes (at a selectable interval) and obtain new zone copies when changes have been made."

Id.

volume DNS data).[49] TSIG serves to verify the origin and authenticity of the DNS data.

TSIG dynamically computes a cryptographic hash in response to a specific DNS request, using the well-known HMAC-MD5 algorithm.

TSIG is felt to be a reasonably mature technology. TSIG depends on a cryptographic signature based on *secret keys*, and thus depends on the sender and the receiver possessing a shared secret. As TSIG does not provide a key distribution mechanism, it would become unwieldy[50] if used to mutually authenticate a large number of systems; however, only a small number of systems typically need to perform (for instance) DNS zone transfers to one another for any particular zone, so TSIG works well enough for its intended purpose.

In comparison with *RFC 2535* DNSSEC, TSIG entails far less computational overhead, and does not increase the size of the DNS database. Lewis describes TSIG as less scalable but more efficient than *RFC 2535* DNSSEC.[51] TSIG provides for authentication and integrity of the data transmitted from the point where it leaves the transmitting server, but it does not authenticate the source data (which may have been compromised in the sending server prior to being transmitted) – in other words, TSIG does not provide full *object security*.[52]

D. Deployment of DNS Security Mechanisms

A number of trial deployments of *RFC 2535* DNSSEC have taken place[53], but on the whole the system is not in production deployment.

In a review undertaken by the IETF in December, 2000, Edward Lewis notes that "[i]n 1999 and 2000, more than a half dozen workshops have been held to test the concepts and the earliest versions of implementations. But to date, DNSSEC is not in common use. The current collective wisdom is that DNSSEC is 1) important, 2) a

[49] PAUL VIXIE ET AL., RFC 2845: SECRET KEY TRANSACTION AUTHENTICATION FOR DNS (TSIG) (Internet Engineering Task Force, May 2000), *at* http://www.ietf.org/rfc.html (updated by *RFC 3645*).

[50] In other words, the two systems participating in a TSIG exchange would have to both know the shared secret through some means other than TSIG itself, since TSIG contains no mechanism for distributing the keys. If the keys are to be transmitted through the Internet, by e-mail for example, they must be protected from disclosure to third parties. All of this adds complexity. Since TSIG is normally used for a bounded set of problems where a trust relationship already exists between two systems, the protocol designers have not felt that this extra complexity was warranted.

[51] *See generally* EDWARD LEWIS, RFC 3130: NOTES FROM THE STATE-OF-THE-TECHNOLOGY: DNSSEC (Internet Engineering Task Force June 2001), *at* http://www.ietf.org/rfc.html.

[52] *See* PAUL VIXIE ET AL., *supra* note 49, at § 6.3; *see also* ATKINS & AUSTEIN, *supra* note 39.

[53] *See* LEWIS, *supra* note 51; *see also* RIPE NCC, *supra* note 45.

buzzword, 3) hard, 4) immature." [54] For *RFC 2535* DNSSEC, this is hardly surprising. As previously noted, the true costs of deployment are high.[55]

In addition, *RFC 2535* DNSSEC appears to suffer from many of the characteristics that, as noted in Section I of this paper, potentially complicate deployment. It is not clear that consumers are willing to pay any premium for DNS security;[56] given that implementation costs (largely in the form of operational complexity) are significant, those who must invest to deploy the technology will find it difficult or impossible to craft a clear business case. *RFC 2535* DNSSEC is strongly influenced by network externality effects – *RFC 2535* DNSSEC would be far more valuable to consumers when it is widely deployed than it is today, or even than it would be if it were in modest production deployment. Moreover, because the system depends on a chain of trust, *RFC 2535* DNSSEC is of limited value until those chains are established all the way from the DNS root to the PC on the consumer's desk without breaks.[57] As all of this implicitly requires the cooperation of many independent parties, the economic transaction costs of a comprehensive deployment would tend to be high.[58]

By contrast, indications are that TSIG is deployable today for zone transfers. Per *RFC 3130*, ". . . one component of DNSSEC, TSIG, is more advanced that the others. Use of TSIG to protect zone transfers is already matured to the 'really good idea to do stage' even if other elements of DNSSEC are not."[59]

Based on the discussion of transaction costs earlier in this paper, this is not surprising. The decision to deploy TSIG concerns only a pair (or a small number) of communicating systems, and in most cases a business relationship already exists between the operators of these systems. Thus, transaction costs to deploy are low, and, as we have seen, ongoing costs for computation and storage are also modest.[60]

[54] LEWIS, *supra* note 51, at § 1.0.

[55] *See supra* Section II.C.1.

[56] There are also open questions regarding the willingness and ability of consumers to cope with the complexity that DNSSEC implies. Suppose the DNSSEC client software were to notify the consumer that the DNS pointer to a commercial web site such as www.amazon.com had been corrupted. It is not clear what action the consumer should then take, since recovery will generally be beyond the consumer's capabilities. In light of this ambiguity, can the DNSSEC client software provide meaningful and sufficient guidance to the consumer?

[57] DNSSEC will be of no use to the average consumer until and unless it is available in the operating system for the consumer's PC – typically Microsoft Windows™.

[58] Some have argued for a more piecemeal, selective approach to deployment, but the DNSSEC standards do not currently embrace this approach.

[59] LEWIS, *supra* note 51.

[60] Unfortunately, the benefits are also modest for the reasons previously noted. The

III. PUBLIC POLICY ALTERNATIVES

To the extent that necessary infrastructure enhancements may not be deployed in the absence of intervention, what is the appropriate role for government?

As we have seen, there is no assurance that industry would deploy a service such as secure DNS based solely on commercial incentives, even assuming the best of intentions on the part of all participants. To the extent that services of this type might be important to the security and robustness of the Internet in the United States, this should be cause for concern.

What role should government play in fostering deployment of Internet capabilities where market forces alone might not suffice? How might government identify and prioritize those capabilities where intervention is warranted (if ever)? For such Internet capabilities as we might deem to be vital, what steps are available to private parties and to the U.S. Government to encourage deployment? Which are likely to be most effective? Which are likely to be least intrusive, and least likely to introduce market distortions?

Most of what we have to say in this section of the paper is not limited to DNS security, and for that matter is not limited solely to cyber security issues. The challenge of promoting the deployment of public goods that provide benefits to the public, but where deployment may not be warranted based solely by the workings of the marketplace, comes up in a great many contexts.

Among the options worth considering by government as a means of fostering deployment of societally valuable services where market incentives might not otherwise suffice are:

1. Provide leadership.
2. Help industry to forge a consensus.
3. Stimulate standards bodies to focus on relevant problems.
4. Collect relevant statistics.
5. Provide "seed money" for research and for interoperability testing.
6. Support desired functionality in products and services through government's own purchasing preferences.
7. Fund the deployment of desired capabilities.
8. Mandate use of desired services.

threats that TSIG guards against are generally irrelevant to the consumer mass market.

An important and overarching consideration is that market intervention should be avoided wherever possible, and kept to a minimum where absolutely necessary. The Communications Act states unambiguously that "[i]t is the policy of the United States . . . to preserve the vibrant and competitive free market that presently exists for the Internet and other interactive computer services, unfettered by Federal or State regulation."[61] Henry David Thoreau stated it more tersely: "That government is best which governs least."[62]

For a somewhat more expansive comment, we turn to a recent study from the Computer Science and Technology Board ("CSTB") of the National Research Council of the National Academies:

> [A]ppropriate market mechanisms could be more successful than direct regulation in improving the security of the nation's IT infrastructure, even though the market has largely failed to provide sufficient incentives for the private sector to take adequate action with respect to information and network security. The challenge for public policy is to ensure that those appropriate market mechanisms develop. How to deal constructively with prevailing market dynamics has been an enduring challenge for government, which has attempted a variety of programs aimed at stimulating supply and demand but which has yet to arrive at an approach with significant impact. Nevertheless, the committee believes that public policy can have an important influence on the environment in which nongovernment organizations live up to their responsibilities for security.[63]

We now discuss the alternative government options in turn, starting with those that are least intrusive.

A. *Provide Leadership*

There may be a tendency to overlook the simplest and least intrusive form by which government can seek to foster change: Simply articulating that change is necessary.

It is perhaps counterintuitive that exercise of "the bully pulpit" alone should be sufficient to influence the behavior of industry participants and

[61] 47 U.S.C. § 230(b)(2) (2000).

[62] HENRY DAVID THOREAU, CIVIL DISOBEDIENCE (1849), *available at* http://www.cs.indiana.edu/statecraft/civ.dis.html (quotation is sometimes attributed to Thomas Jefferson).

[63] INFORMATION TECHNOLOGY FOR COUNTERTERRORISM: IMMEDIATE ACTIONS AND FUTURE POSSIBILITIES 104 (John L. Hennesy et al. eds., 2003) [hereinafter HENNESY ET AL.].

other private citizens,[64] but there is no question that the simple exercise of government leadership has sometimes driven important change.

Leadership in this sense – sometimes referred to as "jawboning" – is more likely to be most effective where some of the following factors hold:

- Government has succeeded in articulating a clear goal that has broad public support.
- The costs associated with doing as the government requests are small (e.g., within the range of discretionary spending of a senior or chief executive).
- The organization that must act needs to curry the favor of the relevant government agency.

B. Help Industry to Forge a Consensus

The U.S. Government frequently provides fora for discussion in order to help industry to reach consensus. The President's Critical Infrastructure Protection Board (CIPB) did so in meeting with the Internet community in the course of preparing the *National Strategy to Secure Cyberspace.*[65]

Analogously, the FCC encourages the communications industry to work together to enhance overall network robustness through the Network Reliability and Interoperability Council (NRIC). NRIC operates under the Federal Advisory Council Act (FACA). As a FACA, the NRIC provides advice to the FCC; further, NRIC often provides guidance regarding best practices to U.S. industry.

In some instances, this consensus could be expressed as a document or guideline prepared by the participants and embodying industry best practices. FACAs often take this approach.

Adhering to industry best practices, as defined by a body such as the NRIC, may also serve to reduce a firm's legal liability to possible allegations of negligence.[66] This form of government participation is

[64] *Cf.* I3P REPORT, *supra* note 21, at 40 ("Currently, the federal government's approach relies on public-private partnerships *and the influence of persuasion*; more rigorous analysis needs to be done on the prospects for success of this approach.") (emphasis added).

[65] DRAFT NATIONAL STRATEGY TO SECURE CYBERSPACE, *supra* note 3.

[66] Potential tort liability, where a firm might be alleged to have taken less than reasonable care to secure its infrastructure against cyberattacks is an emerging, but still largely undeveloped area of the law. *See* CRITICAL INFORMATION INFRASTRUCTURE PROTECTION AND THE LAW: AN OVERVIEW OF KEY ISSUES (Cynthia A. Patterson & Stewart D. Personick eds., 2003), *available at* http://www7.nationalacademies.org/cstb/pub_ciip.html [hereinafter CRITICAL INFORMATION INFRASTRUCTURE PROTECTION AND THE LAW].

generally viewed as positive by industry and by the broader community. It provides government with the opportunity to offer leadership in a minimally intrusive way.

This form of government participation provides industry with an additional benefit. Companies that routinely compete in the marketplace are understandably uncomfortable meeting to discuss joint action, for fear that their discussions could be misconstrued as being anticompetitive. To the extent that the U.S. Government calls firms together to discuss specific issues in the public interest, antitrust concerns tend to be mitigated.[67]

C. Stimulate Standards Bodies to Focus on Relevant Problems

One form of industry consensus is embodied in the standards process. As described above, government could play a role in helping industry to agree on a standard. If appropriate, government could perhaps reinforce this result by encouraging the relevant standards body or bodies to officially adopt a standard reflecting that consensus.

In general, government would look to industry to develop solutions for the standards process. Government is not well equipped to pick winners and losers.

For some standards bodies, notably including the International Telecommunications Union (ITU), formal U.S. Government advocacy can play a crucial role in achieving adoption of a standard.

The Internet Engineering Task Force (IETF) is the primary standards body for the Internet. By long-standing tradition, the IETF expects standards participants to present their views as an individual expert, rather than those of the organizations that they represent. The U.S. Government thus plays no formal role in the IETF. Even in this case, however, government can when appropriate facilitate the standards process by supporting research and interoperability testing and by identifying problem areas where it appears that the public interest would be well served by a standards-based solution.

[67] As a somewhat related example, the *National Strategy to Secure Cyberspace* recognizes the importance of establishing mutual assistance agreements to help infrastructure sectors respond to cybersecurity emergencies. *See* NATIONAL STRATEGY TO SECURE CYBERSPACE, *supra* note 4, at 24 (stating that the "[Department of Justice] and the Federal Trade Commission should work with the sectors to address barriers to such cooperation, as appropriate." (emphasis omitted)).

D. Collect Relevant Statistics

In a competitive communications industry, industry participants will have data about their own experiences, but no single industry participant will necessarily have a global view.[68]

Government can collect data where appropriate to identify problems, to determine their magnitude, and to provide a basis on which to evaluate potential solutions.

In determining whether to do so, it would invariably be necessary to balance several conflicting objectives. There may be compelling public interest reasons for gathering certain kinds of information; however, collecting that information represents a regulatory burden on the companies involved. That burden should be avoided where possible, and minimized where the data are truly needed.

Another tension of objectives relates to the sensitivity of data gathered. The public has a right to know information held by the Government, as embodied in the Freedom of Information Act (FOIA) and also by various state "sunshine" acts. At the same time industry participants have a legitimate interest in protecting competitively sensitive information, and in preserving the privacy of their customers. Often, these conflicting demands have been reconciled by having a third party anonymize data before providing it to the Government.[69]

There are specific exemptions from FOIA that address specific needs. One recent report rightly observes that these exemptions provide agencies with substantial ability to shield information of this type from inappropriate disclosure under FOIA;[70] however, that knowledge offers little comfort to industry participants, who must consider not only whether government *can* avoid inappropriate disclosure of their sensitive data, but also whether it *will*.[71]

[68] *Cf.* NATIONAL STRATEGY TO SECURE CYBERSPACE, *supra* note 4, at 19 ("There is no synoptic or holistic view of cyberspace. Therefore, there is no panoramic vantage point from which we can see attacks coming or spreading.").

[69] For example, when industry participants provide incident reports to Information Sharing and Analysis Centers (ISACs) operating under PDD-63, the information might be sanitized or anonymized before being shared with other ISAC participants or with the government.

[70] *See* CRITICAL INFORMATION INFRASTRUCTURE PROTECTION AND THE LAW, *supra* note 66, at 25-29.

[71] Notably, the Homeland Security Act specifically exempts information about critical infrastructure vulnerabilities provided voluntarily from FOIA obligations. *Cf.* PRESIDENT'S CRITICAL INFRASTRUCTURE PROTECTION BOARD, *supra* note 4, at 25 ("the legislation encourages industry to share information with DHS by ensuring that such voluntarily provided data about threats and vulnerabilities will not be disclosed in a manner that could damage the submitter." This is an area of ongoing concern for the DHS, which is working to ". . .

In those instances where data collection appears warranted in support of some public policy objective, government can work with industry to define the data required, to evaluate necessary safeguards on the dissemination of that information, and then to establish voluntary reporting programs.

Mandatory reporting can be appropriate in some circumstances, but only where the need for the data is compelling, where the data to be collected is well and narrowly defined, and where voluntary reporting for some reason is either inappropriate or unsuccessful.

E. Provide "Seed Money" for Research and for Interoperability Testing

For facilities that may benefit the public interest, but not necessarily individual users or industry participants, it may be that no private funding source is motivated to provide initial "seed" money. Certain security services, for instance, may benefit the public at large rather than any particular individual or company.

Public funding (or funding by public interest sources) may be the only practical way to foster development of such capabilities.

Analogous issues exist with interoperability testing. Many network services are useful only to the extent that they are interoperable with their counterparts in other networks. These counterpart services may be implemented independently and in competing products. Absent testing, there is no assurance that these implementations will interoperate correctly.

The government role in such activities is well established and widely accepted. For an example where this approach worked brilliantly, see the discussion of "Funding for the early Internet – a happier case study" later in this paper. Research[72] and interoperability testing may, in addition, serve to facilitate the standards process. The IETF will not progress a standard to Draft Standard status until interoperability among independent implementations has been rigorously demonstrated.[73]

establish uniform procedures for the receipt, care, and storage . . . of critical infrastructure information that is voluntarily submitted to the government.").

[72] *See* PRESIDENT'S CRITICAL INFRASTRUCTURE PROTECTION BOARD*, supra* note 4, at 34-35 (explicitly recognizing the importance of prioritizing the Federal research and development agenda and tasking the OSTP with doing so).

[73] BRADNER, *supra* note 44.

F. Support Desired Functionality in Products and Services Through Government's Own Purchasing Preferences

To the extent that the U.S. Government is itself a significant user of data networking services, its buying preferences for its own use can serve to influence the evolution of technology.

This represents an interesting proactive lever for change. Industry and the public tend to view this mechanism as legitimate and non-intrusive. It alters the economic incentives of suppliers, but it works *with* the economic system rather than against it.

This form of intervention may be particularly useful as a means of motivating suppliers (e.g., of software) to include desired functionality with the standard distribution versions of their products.

At the same time, it should not be viewed as a panacea. Government purchasing power may not be sufficient to drive widespread adoption (which is still subject to the economic effects of network externalities of the larger market).[74] Consequently, there is always the risk that government will pay a substantial premium in a vain attempt to foster the development and deployment of features and services that, at the end of the day, prove to be of limited utility.

A case in point is the U.S. Government OSI Profile (GOSIP). A massive international standardization effort was in play in the Eighties and into the Nineties on the part of the International Organization for Standardization (ISO) and the Telecommunication Standardization arm of the International Telecommunications Union (ITU-T).[75] They were seeking to develop an entire family of data communications protocols, based on principles of *Open Systems Interconnection* (OSI). The OSI protocols reflected modern concepts of protocol layering, and a full set of applications, including virtual terminal, file transfer, electronic mail, directory, and network management.

It might seem odd in retrospect that the global standards bodies and governments set out to recreate out of whole cloth functionality that already existed. OSI was nominally open to multiple vendors and implementations, but no more so than TCP/IP. Indeed, at the end of

[74] *Cf.* HENNESSY ET AL., *supra* note 63, at 103 ("the IT sector is one over which the federal government has little leverage. IT sales to the government are a small fraction of the IT sector's overall revenue, and because IT purchasers are generally unwilling to acquire security features at the expense of performance or ease of use, IT vendors have little incentive to include security features at the behest of government alone.").

[75] At the time, this was the International Telephone and Telegraph Consultative Committee (CCITT). *See* INTERNATIONAL TELECOMMUNICATIONS UNION, ITU OVERVIEW – HISTORY (Feb. 13, 2002), *at* http://www.itu.int/aboutitu/overview/history.html.

the day, OSI provided no new functionality that users found significant that was not already available under the TCP/IP protocol suite.

Many foreign governments considered TCP/IP to be the creation of the U.S. Department of Defense. Because TCP/IP had not been created by the recognized international standards process, they considered it inappropriate as the basis for a new, global family of communications standards.

The U.S. Government attempted to join a global bandwagon forming in favor of OSI. The National Institutes for Standards and Technology (NIST) published GOSIP Version 1[76] in August 1988, and followed a year later with GOSIP Version 2.[77] A profile was needed because many of the OSI protocols were so specified as to permit a variety of mutually incompatible possible realizations.[78] As of August 1990, Federal agencies were required to acquire OSI products when they required the functionality supplied by the OSI features specified in GOSIP. There was, however, no requirement that Federal agencies procure *only* GOSIP-compliant implementations for these purposes, nor was there an obligation for Federal agencies to *use* the GOSIP-compliant implementations that they had thus procured.

OSI protocols had developed what might have seemed to be an unbreakable momentum in the late Eighties. The ISO and CCITT unequivocally backed the protocols, while the Internet standards groups accepted at least an extended period of coexistence between TCP/IP and OSI protocols.[79] Digital Equipment Corporation (DEC), at the time a leading computer manufacturer, had committed to implementing OSI communications protocols in DECNET Phase V.

Today, however, OSI protocols serve as little more than a historical curiosity, an interesting footnote. Why is it that OSI protocols failed to achieve broad market acceptance?

Some have argued (and sometimes with surprising vehemence) that government support was the kiss of death for OSI protocols. This seems, however, to miss the point. In particular, it fails to explain the

[76] Approval of Federal Information Processing Standards Publication 146, Government Open Systems Interconnection Profile (GOSIP), 53 Fed. Reg. 32,270, 32,270-02 (Dep't Commerce Aug. 24, 1988).

[77] Proposed Revision of Federal Information Processing Standard (FIPS) 146, G3OSIP, 54 Fed. Reg. 29,597, 29,597-602 (Dep't Commerce July 13, 1989).

[78] There was no assurance that two independent implementations of, say, the FTAM file transfer and access method would interoperate correctly. This is much less of an issue for TCP/IP protocols, where demonstrated interoperability is a prerequisite to standardization. It would be unusual, for instance, for the FTP support in two different TCP/IP implementations to fail to interoperate correctly.

[79] *See* V. CERF & K. MILLS, RFC 1169: EXPLAINING THE ROLE OF GOSIP (Internet Engineering Task Force, Aug. 1990), *at* http://www.ietf.org/rfc.html.

success of TCP/IP protocols, which by all accounts benefited enormously from substantial support from the U.S. Government.

Others have argued that OSI protocols were cumbersome, and evolved slowly, because they were developed by large committees and because the protocol specification effort took place *in advance of* implementation. (Internet protocols, by contrast, would never be standardized until independent implementations had been shown to interoperate.) There probably is some truth to this assertion, and it is moreover plausible in terms of what we know of the economics of transaction costs – the need to obtain concurrence of a great many independent parties invariably exacts costs, one way or another. Nonetheless, it is only a part of the answer.

It must also be noted that OSI protocol implementations tended to be significantly more expensive than TCP/IP protocol implementations, not only in terms of purchase price, but also in terms of memory requirements, processing power requirements, and operational complexity. These were certainly factors, but they may not have been decisive.

A simple and sufficient explanation flows from the economic theory of network externalities. TCP/IP implementations were available on most platforms of interest, and the software was inexpensive or free in many cases, unlike OSI implementations. The deployment of OSI protocols at their peak probably never accounted for more than 1-2% of all traffic on the Internet. Users were motivated to use TCP/IP, because most of the content that they wanted to use or view was available in the TCP/IP world, and not in the OSI world. Content providers and application developers were motivated to use TCP/IP, because the majority of their prospective users were TCP/IP users. (Similar factors may have provided Microsoft Windows with an advantage over the Macintosh and, for that matter, VHS with an advantage over Beta, as noted earlier.)

OSI protocols were starting from a position of zero market share. They could not fully supplant TCP/IP protocols unless they replaced *all* of TCP/IP's functionality; however, TCP/IP began with a huge head start in functionality. Moreover, ongoing investment in new functionality based on the TCP/IP protocols inevitably outstripped that for new OSI functionality by a wide margin. Given that OSI had no compelling inherent advantage over TCP/IP, there was never any means to reverse this trend.

Eventually, the requirement to procure services implementing GOSIP (and its companion standard, the Government Network

Management Profile (GNMP))[80] was lifted. It was presumably recognized that a mandate to procure GOSIP-compliant solutions no longer served a useful purpose. Meanwhile, the U.S. Government had supported the evolution and testing of OSI protocols in many ways, and Federal agencies likely paid more than they otherwise might have to procure functionality that they ultimately did not need and, for the most part, did not use.

G. Fund the Deployment of Desired Capabilities

If deployment of a service is in the public interest, but not in the individual interest of the firms that must deploy it, and if deployment entails significant costs, then those firms have a significant economic disincentive to deploy. In a competitive, deregulated telecommunications marketplace, it is not clear how those firms could recapture their investment.

In those cases, it may be that the only possibility of achieving widespread deployment will be through some combination of subsidizing or funding that deployment as well as any associated incremental operational costs, or possibly by mandating deployment, or both.

The Communications Assistance for Law Enforcement Act (CALEA) is a case in point.[81] CALEA establishes carrier obligations in regard to lawful intercept of communications (e.g. wiretap). No telecommunications customer would wish to pay a premium for the privilege of having his or her own communications amenable to wiretap, nor would any carrier have a business incentive to implement the necessary tools and facilities.

As a result, CALEA establishes the Department of Justice Telecommunications Carrier Compliance Fund[82] in an effort to "make the carriers whole." This process has not been painless – carriers have argued that the fund does not adequately reimburse them for costs incurred.[83]

[80] Approval of Federal Information Processing Standards Publications (FIPS) 146-2, Profiles for Open Systems Internetworking Technologies; and 179-1, Government Network Management Profile, 60 Fed. Reg. 25,888-02 (Nat'l Inst. of Standards and Tech. May 15, 1995), *available at* http://www.itl.nist.gov/fipspubs/fip179-1.htm.

[81] Communications Assistance for Law Enforcement Act, Pub. L. No. 103-414, 108 Stat. 4279 (1994) (codified as amended in scattered sections of 18 U.S.C. and 47 U.S.C.) . For a brief background on CALEA, see FCC, CALEA, *at* http://www.fcc.gov/calea/ (last reviewed/updated 6/10/04).

[82] Communications Assistance for Law Enforcement Act § 401 (codified as amended at 47 U.S.C. § 1021 (2000)).

[83] In practice, the fund reimburses equipment suppliers. There has been to the author's knowledge only one instance where the fund was used to reimburse a service provider. Service providers incur costs for software upgrades to deploy CALEA, and they incur significant additional deployment costs beyond those associated with hardware and software.

Government funding for public goods can take any of a number of forms. It can come from general revenues. It can be a distinct fund, as is the case for CALEA. It can also be a separate fund privately managed on behalf of the government, as is the case for universal service.

H. Mandate Use of Desired Services

If functionality were truly deemed to be essential to the public interest, and if market forces were insufficient to ensure its deployment, then it could in principle be appropriate for government to mandate its deployment and use.

For the Internet, there is no obvious historical example; however, there are many examples in the history of the telephone industry in the United States.

One of these is the previously-noted CALEA. CALEA serves both to oblige telecommunications carriers to provide the technical means of achieving lawful intercept (wiretap) and to provide a mechanism for offsetting their costs in doing so. Lawful intercept is a legitimate societal need, but it does not specifically benefit an individual carrier; consequently, it can only be achieved to the extent that government provides the impetus, in this case by means of an explicit mandate.

Other examples of services that might have been unlikely to deploy absent government action include:

- Disabilities access to telecommunications,[84]
- Provision of 911 services, and
- Local number portability.[85]

This is the most intrusive means the government has of driving deployment. For a number of reasons, it should be used sparingly.[86]

First, as our experience with GOSIP demonstrates, government's ability to prognosticate is limited.[87] If government is to mandate deployment and use, it must be very certain that the functionality in question is truly necessary.

[84] 47 U.S.C. §§ 225, 255 (2000).

[85] *Id.* at § 251.

[86] *Cf.* I3P REPORT , *supra* note 21, at 41 ("Aggressive approaches that more fully use the powers of the federal and state governments are also possible, but the costs and benefits are not well understood and the reasons for a general reluctance to regulate are well known. This statement raises the question of who is responsible for security in this information infrastructure 'commons' and who should pay for it.").

[87] *Cf.* HENNESSY ET AL., *supra* note 63, at 103-104 ("it is likely that attempts at such regulation will be fought vigorously, or may fail, because of the likely inability of a regulatory process to keep pace with rapid changes in technology.").

Second, mandating a function will generally have a tendency to distort the relevant market. Wherever possible, market mechanisms should be preferred over mandates, especially unfunded mandates.

Finally, there is the risk that a government mandate might lock the industry into the use of a particular technology long after market forces would otherwise have obsoleted it.

I. Adoption of the Metric System – A Sobering Case Study

In considering the prospects for achieving deployment by means of government actions short of an outright mandate, it is helpful to consider historical precedents. We have already discussed GOSIP. Another example, albeit from a different technological domain, is conversion to the metric system.

In 1971, the National Bureau of Standards published a report, *A Metric America*,[88] recommending "[t]hat the Congress, after deciding on a plan for the nation, establish a target date ten years ahead, by which time the U.S. will have become predominantly, though not exclusively, metric. . . ."[89]

The benefits of metric conversion were thought to be manifest. Recognizing this, the U.S. Government has undertaken significant efforts over the years to foster adoption of the metric system,[90] including the passage of the Metric Conversion Act of 1975[91] and the issuance of Executive Order 12770[92] in 1991. Nonetheless, thirty-two years after the publication of *A Metric America*, it can hardly be said that the United States has "become predominantly, though not exclusively, metric".

In *A Metric America*, the National Bureau of Standards report recognized that the United States had become an isolated island in a metric world, and identified the potential costs associated with that isolation. They also attempted to quantify the costs of conversion, and the potential benefits – largely in terms of global trade and simplified

[88] NAT'L BUREAU OF STANDARDS, A METRIC AMERICA: A DECISION WHOSE TIME HAS COME*,* NBS Special Publication 345, July 1971.

[89] *Id.* at iii.

[90] Interest in the metric system in the U.S. actually began much earlier. John Quincy Adams considered it in his *Report Upon Weights and Measures* in 1821. JOHN QUINCY ADAMS, REPORT ON WEIGHTS AND MEASURES (1821). Beginning in 1866, a series of laws were enacted that legalized the use of metric weights and measures, and directed the Postmaster General to distribute metric postal scales to all post offices exchanging mail with foreign countries. *See* NAT'L BUREAU OF STANDARDS, *supra* note 88. In fact, the U.S. became the first officially metric country by adopting the metric standards in the *Treaty of the Meter* to be the nation's "fundamental standards" of weight and mass in 1889. *Id.* at 14-15.

[91] Metric Conversion Act, Pub. L. No. 94-168, 89 Stat. 1007 (1975) (codified as amended in 15 U.S.C. § 205 (2000)).

[92] Exec. Order No. 12,770, 50 Fed. Reg. 35,801 (July 25, 1991), *available at* http://ts.nist.gov/ts/htdocs/200/202/pub814.htm#president.

education. The Metric Conversion Act of 1975 expressed the advantages in unambiguous bread and butter terms:

> (3) World trade is increasingly geared towards the metric system of measurement.
> (4) Industry in the United States is often at a competitive disadvantage when dealing in international markets because of its nonstandard measurement system, and is sometimes excluded when it is unable to deliver goods which are measured in metric terms.
> (5) The inherent simplicity of the metric system of measurement and standardization of weights and measures has led to major cost savings in certain industries which have converted to that system.
> (6) The Federal Government has a responsibility to develop procedures and techniques to assist industry, especially small business, as it voluntarily converts to the metric system of measurement.
> (7) The metric system of measurement can provide substantial advantages to the Federal Government in its own operations.[93]

An important collective effect of the Metric Conversion Act and of Executive Order 12770 has been to require that each Federal agency ". . . to the extent economically feasible by the end of the fiscal year 1992, use the metric system of measurement in its procurements, grants, and other business-related activities, except to the extent that such use is impractical or is likely to cause significant inefficiencies or loss of markets to United States firms, such as when foreign competitors are producing competing products in non-metric units."

The Metric Conversion Act also attempts to "seek out ways to increase understanding of the metric system of measurement through educational information and guidance and in Government publications." The Act established a United States Metric Board[94] tasked with carrying out "a broad program of planning, coordination, and public education." The Board was to perform extensive public outreach, to "encourage activities of standards organizations," to liaise with foreign governments, to conduct research and surveys, to "collect, analyze, and publish information about the usage of metric measurements," and to "evaluate the costs and benefits of metric usage." Thus, the metric conversion program attempted, to a lesser or greater degree, to employ essentially every tool available to government short of outright deployment funding or an explicit mandate.[95]

[93] Metric Conversion Act, 89 Stat. 1007.
[94] *Id.*
[95] *Id.*

These efforts undoubtedly had effect, but not as great an effect as was intended. Why was this?

> A variety of reasons have been put forward to explain why the metric transition has not made widespread progress in the U.S. in the past. They include lack of national leadership, reluctance to embark on such a change, and *the failure of the voluntary effort that began in 1975*. The many competing national priorities and *the lack of immediate and visible benefit to a transition* clearly were factors. There are political, economic, and social reasons to explain the apparent slow progress and reluctance to make the transition.[96]

It is not the intent of this paper to trivialize or over-simplify what undoubtedly was a very complex process. The key point that the reader should take away from this case study is that, for certain kinds of innovations where economic incentives are not sufficient to motivate their deployment in a free market system, there can be no assurance that government actions short of deployment funding or an explicit mandate will generate substantial deployment.

J. Funding for the Early Internet – A Happier Case Study

In the case of the Internet, by contrast, the historic effects of direct Government funding have in most instances been salutary. The original ARPAnet, the predecessor to the Internet, was funded in the late Sixties by the Advanced Research Projects Agency of the U.S. Department of Defense (DARPA).[97]

In the early Eighties, DARPA funded the University of California at Berkeley to incorporate TCP/IP protocols into Berkeley UNIX®.[98] This effort produced one of the most widely used TCP/IP implementations. Berkeley UNIX was incorporated into an emerging generation of UNIX workstations, thus fostering precisely the network externalities effects that ultimately enabled TCP/IP to prevail in the marketplace.

[96] DR. GARY P. CARVER, NAT'L INST. OF STANDARDS & TECH., A Metric America: A Decision Whose Time Has Come – For Real, NISTIR 4858 (1992), *available at* http://ts.nist.gov/ts/htdocs/200/202/4858.htm (emphasis added). Dr. Carver was then chief of the Metric Program at the National Institutes of Standards and Technology (NIST).

[97] BARRY M. LEINER ET AL, INTERNET SOCIETY, A BRIEF HISTORY OF THE INTERNET (Dec. 10, 2003), *at* http://www.isoc.org/internet/history/brief.shtml#Origins. Note that the Advanced Research Projects Agency (ARPA) changed its name to Defense Advanced Research Projects Agency (DARPA) in 1971, then back to ARPA in 1993, and back to DARPA in 1996.

[98] *Id.*

The U.S. National Science Foundation (NSF) provided initial funding for CSNET as a limited-function network for the academic research community. The NSF then invested an estimated $200 million from 1986 to 1995 to build and operate the NSFNET as a general purpose Internet backbone for the research and education community.[99]

Most observers would agree that the modest investments that DARPA and the NSF made in the Internet have collectively been a brilliant success.

IV. CONCLUDING REMARKS

On a hasty reading, this paper might be construed as advocating that government take an intemperate, interventionist approach toward the Internet.

What is called for, in the author's view, is a reasoned and balanced approach. Much has been made of the lack of regulation of the Internet.[100] Yet the very existence of the Internet is a direct result of a succession of government interventions, many of them highly successful. Among these were the initial funding of the ARPAnet, the FCC's Computer Inquiries (simultaneously deregulating services like the Internet while opening up underlying telecommunications facilities for their use), support for CSNET and the NSFNET, and the funding of TCP/IP protocol implementation in Berkeley UNIX.[101] Each of these achieved important and positive results without resorting to a regulatory mandate.

There have also been failures of government intervention. Perhaps the most relevant was the U.S. Government's support of OSI protocols through GOSIP and the GNMP, as described earlier in this paper. That ultimately unsuccessful attempt to use the purchasing power of government to promote global standards that the marketplace had by and large not demanded, likely resulted in significant diversion of attention and waste of resources on the part of both government and industry.

Another example was metric conversion, where the U.S. Government has attempted a combination of practically every conceivable measure short of an outright mandate but has not achieved the widespread deployment that was hoped for.

[99] *Id.*

[100] *See* JASON OXMAN, THE FCC AND THE UNREGULATION OF THE INTERNET (FCC Office of Plans and Policy, Working Paper No. 31, July 1999), *available at* http://ftp.fcc.gov/Bureaus/OPP/working_papers/oppwp31.pdf.

[101] LEINER ET AL., *supra* note 97.

Government is neither omniscient nor omnipotent. Government could do too little. Government could also do too much. How to know which is which?

Two principles may be useful going forward:

> BALANCE: Government should recognize both the risks of action and those of inaction, and make cautious and deliberate choices.

> MINIMALISM: Government should choose to err in general on the side of less regulation rather than more. Do not attempt a massive intervention where a less intrusive intervention might suffice. Do not intervene at all unless markets have shown themselves to be unable to deliver a socially important outcome.

A MODEL FOR WHEN DISCLOSURE HELPS SECURITY:

WHAT IS DIFFERENT ABOUT COMPUTER AND NETWORK SECURITY?

PETER P. SWIRE[*]

TABLE OF CONTENTS

[*] Professor of Law and John Glenn Research Scholar in Public Policy Research, Moritz College of Law of the Ohio State University. Many people from diverse disciplines have contributed to my thinking on the topic of this paper. An earlier version of the project, entitled "What Should Be Hidden and Open in Computer Security: Lessons from Deception, the Art of War, Law, and Economic Theory," was presented in 2001 at the Telecommunications Policy Research Conference, the Brookings Institution, and the George Washington University Law School, *available at* www.peterswire.net. I am also grateful for comments from participants more recently when this topic was presented at the Stanford Law School, the University of North Carolina Law School, and the Silicon Flatirons Conference at the University of Colorado School of Law. Earlier phases of the research were supported by the Moritz College of Law and the George Washington University Law School. Current research is funded by an award from the John Glenn Institute for Public Service & Public Policy.

INTRODUCTION

This article asks the question: "When does disclosure actually help security?" The question of optimal openness has become newly important as the Internet and related technologies have made it seem inevitable that information will leak out. Sun Microsystems CEO Scott McNealy received considerable press attention a few years ago when he said: "You have zero privacy. Get over it."[1] An equivalent statement for security would be to say: "You have zero secrecy. Get over it." Although there is a germ of truth in both statements, neither privacy nor secrecy is

[1] A. Michael Froomkin, *The Death of Privacy*, 52 STAN. L. REV. 1461, 1462 (2000).

or should be dead. Instead, this article seeks to provide a more thorough theoretical basis for assessing how disclosure of information will affect security. In particular, this article seeks to understand what is different between traditional security practices in the physical world, on the one hand, and best practices for computer and network security, on the other.

The discussion begins with a paradox. Most experts in computer and network security are familiar with the slogan that "there is no security through obscurity."[2] For proponents of Open Source software,[3] revealing the details of the system will actually tend to improve security, notably due to peer review. On this view, trying to hide the details of the system will tend to harm security because attackers will learn about vulnerabilities, but defenders will not know where to patch the vulnerabilities. In sharp contrast, a famous World War II slogan says "loose lips sink ships."[4] Most experts in the military and intelligence areas believe that secrecy is a critical tool for maintaining security.

Section I of this article provides a basic model for deciding when the Open Source and military/intelligence viewpoints are likely to be correct. Insights come from a 2x2 matrix. The first variable is the extent

[2] A search on Google for "security obscurity" discovered 110,000 web sites with those terms. Reading through the web sites show that a great many of them discuss some version of "there is no security through obscurity."

[3] Wikipedia, an on-line encyclopedia that uses Open Source approaches, defines "open source" as:

> "a work methodology that fits the Open Source Definition, and generally is any computer software whose source code is either in the public domain or, more commonly, is copyrighted by one or more persons/entities and distributed under an open-source license such as the GNU General Public License (GPL). Such a license may require that the source code be distributed along with the software, and that the source code be freely modifiable, with at most minor restrictions."

WIKIPEDIA, OPEN SOURCE, *at* http://en.wikipedia.org/wiki/Open_source (last modified Aug. 5, 2004). Source code is defined as:

> "Source code (commonly just source or code) refers to any series of statements written in some human-readable computer programming language. In modern programming languages, the source code which constitutes a software program is usually in several text files, but the same source code may be printed in a book or recorded on tape (usually without a filesystem). The term is typically used in the context of a particular piece of computer software. A computer program's *source code* is the collection of files that can be converted from human-readable form to an equivalent computer-executable form. The source code is either converted into executable by an software development tool for a particular computer architecture, or executed from the human readable form with the aid of an interpreter."

WIKIPEDIA, SOURCE CODE, *at* http://en.wikipedia.org/wiki/Source_code (last modified Aug. 5, 2004).

[4] For images of World War II posters on the subject, see New Hampshire State Library, *Unifying a Nation*, *available at* http://www.state.nh.us/ww2/loose.html. The posters tell vivid stories. One poster has a picture of a woman and the words "Wanted for Murder: Her Careless Talk Costs Lives." Another shows a sailor carrying his kit, with the words "If You Tell Where He's Going . . . He May Never Get There."

to which disclosure is likely to help the attackers, by tipping off a vulnerability the attackers would otherwise not have seen. The second variable is the extent to which the disclosure is likely to improve the defense. Disclosure might help the defense, notably, by teaching defenders how to fix a vulnerability and by alerting more defenders to the problem. The 2x2 matrix shows the interplay of the help-the-attacker effect and the help-the-defender effect, identifying four basic paradigms for the effects of disclosure on security: the Open Source paradigm; the Military/Intelligence paradigm; the Information Sharing paradigm; and the Public Domain.

Section II provides an explanation of why many computer and network security issues are different from military and other traditional security problems of the physical world. The discussion focuses on the nature of the "first-time attack" or the degree of what the paper calls "uniqueness" in the defense. Many defensive tricks, including secrecy, are more effective the first time there is an attack on a physical base or computer system. Secrecy is far less effective, however, if the attackers can probe the defenses repeatedly and learn from those probes. It turns out that many of the key areas of computer security involve circumstances where there can be repeated, low-cost attacks. For instance, firewalls, mass-market software, and encryption systems all can be attacked repeatedly by hackers. Under such circumstances, a strategy of secrecy – of "security through obscurity" – is less likely to be effective than for the military case.

Even recognizing the lower effectiveness of secrecy in many computer and network applications, there will still often be advantages of secrecy in practice. Section III relaxes the assumptions of the model presented in Section I. The Open Source approach makes three assumptions: (1) disclosure will offer little or no help to attackers; (2) disclosure will tend to upgrade the design of defenses; and (3) disclosure will spread effective defenses to third parties. In practice, secrecy will often be of greater use than the Open Source advocates have stated, because one or more of the three assumptions will not hold. Section III explains some of the major categories of situations where secrecy is likely to be more or less effective at promoting security.

The chief intellectual task of this article is to help us think about when disclosure will help or harm security. There are other major considerations that go into an informed judgment about whether to disclose information about a security vulnerability. For instance, it may promote accountability and the long-run health of the system to err on the side of disclosure. This instinct underlies the Freedom of

Information Act[5] and many other laws and practices encouraging disclosure. As another example, disclosure can compromise personal privacy in some circumstances. Accountability and privacy are vital goals in the overall analysis of when to disclose information. Discussion of those goals figures prominently in my larger research project on openness and security. This article, however, focuses on when disclosure will help the specific goal of system security: when will disclosure protect against the attacker gaining control of a physical installation or computer system.

I. A MODEL FOR WHEN DISCLOSURE HELPS SECURITY

When does disclosure help security? The intuition for experts in the military and intelligence realms is usually that secrecy (the lack of disclosure) is an essential tool for enhancing security. Military bases and weapon systems are cloaked in secrecy. Intelligence agencies tell little about their capabilities, sources, and methods. The slogan for this position is the World War II motto that "loose lips sink ships." The graphic image is that too much disclosure ("'loose lips") will tip off the enemy where to send its submarines ("sink ships").[6] In such instances, disclosure can be tantamount to treason.

A. Case A: The Open Source Paradigm

Despite the World War II intuition, a pervasive theme of many computer security discussions is that "there is no security through obscurity."[7] For people outside of the computer security realm, it may initially be difficult to understand how that slogan has become a truism. Based on research and discussions with computer security researchers, there seem to be three assumptions-often implicit-that under-gird the slogan.

[5] *See* 5 U.S.C. § 552, *amended by* Pub. L. No. 104-231, 110 Stat. 3048 (1996).

[6] *See* New Hampshire State Library, *supra* note 4.

[7] *Supra* note 2. The origin of the slogan "there is no security through obscurity" is obscure. I would welcome information on the origins of the term. It was certainly used by the early 1990's. *See, e.g.*, *Netware Users React to Security Threat*, INTERNET WEEK (Oct. 5, 1992) (Rop Gonggrijp refers to "security through obscurity" as a policy used by Novell).

In considering whether to disclose a vulnerability, supporters of openness seem to assume the following:

(A1) Attackers will learn little or nothing from the disclosure.

(A2) Disclosure will prompt the designers to improve the design of defenses.

(A3) Disclosure will prompt other defenders to take action.

The discussion below in Sections II and III will develop in more detail the intuitions that underlie these three assumptions. It will also critically examine each assumption. For the present, however, the basic idea for assumption (A1) is that software and network vulnerabilities, once discovered by any attacker, will often quickly become known to other attackers. For instance "warez" sites and other mechanisms exist to teach hackers about new attacks.[8] Public disclosure of a vulnerability will thus not significantly help attackers exploit the vulnerability.

The basic idea for assumption (A2) is a deeply-held tenet in the Open Source movement. The idea is that software will improve quickly if a wide array of programmers can see the code, find flaws in it, and fix those flaws. In the words of researchers Randy Bush and Steven Bellovin: "Hiding security vulnerabilities in algorithms, software, and/or hardware decreases the likelihood they will be repaired."[9]

The basic idea for assumption (A3) is that many people may be affected by a vulnerability other than the software or system designers. For a software program, for instance, assumption (A2) is directed at the group of programmers who may write new code to improve the software. There are likely many system owners, however, who use the software program but are not involved in writing it. Assumption (A3) focuses on how disclosure of a vulnerability can improve the security of these system owners. System owners who learn of the vulnerability can install a patch[10] or upgrade once it is available. If a patch is not yet available, the

[8] *E.g.*, http://easywarez.com; http://ICEWAREZ.net (examples of "warez" sites that provide downloads of software illegally, including software that can be used for hacking purposes).

[9] *E.g.*, RANDY BUSH & STEVEN M. BELLOVIN, RFC 2026: SECURITY THROUGH OBSCURITY DANGEROUS (Internet Eng'g Task Force, Working Paper, Aug. 21, 2002), *at* http://www.research.att.com/~smb/papers/draft-ymbk-obscurity-00.txt.

[10] *See* WIKIPEDIA, PATCH, *at* http://webopedia.internet.com/TERM/p/patch.html (last modified Aug. 5, 2004) (defining "patch:" "Also called a service patch, a fix to a program bug. A patch is an actual piece of object code that is inserted into (patched into) an executable program. Patches typically are available as downloads over the Internet." *See also Understanding Patch and Update Management: Microsoft's Software Update Strategy*, Oct. 1, 2003, *at* http://www.microsoft.com/technet/security/topics/patch/patchmanagement.mspx.

system owner can decide to take other measures, such as taking a system off-line or disabling the software, until a defense does become available.

Effects of assumptions (A1), (A2), and (A3): in the Open Source paradigm, the costs of disclosure of a vulnerability are low because attackers learn little or nothing from the disclosure. The benefits of disclosure are high because of improved system design and actions taken by non-designers to protect their systems.

B. Case B: The Military Paradigm

The assumptions in the military setting are directly contrary to the Open Source paradigm:

(B1) Attackers will learn a lot from disclosure of a vulnerability.

(B2) Disclosure will teach the designers little or nothing about how to improve the defenses.

(B3) Disclosure will prompt little or no improvement in defense by other defenders.

The intuition for assumption (B1) is that it is difficult in a military setting for the attackers to learn about a vulnerability. Consider a hill that is being defended by mines or camouflaged machine guns. Should the defenders publish the location of the defenses on the Internet? The answer clearly is no. It will be difficult and costly for the attackers to learn those locations and to determine the least-defended path up the hill. Colloquially and literally, the attackers will have to "pay in blood" to learn the weak points of the defense. Disclosure in this setting would help attackers considerably.

The intuition for assumption (B2) is a bit less clear-cut. It certainly is possible that public disclosure of a design will lead clever persons outside of the military to suggest improvements in design. More likely, however, the incremental learning from these outsiders will be modest at best. For specialized military topics, there is likely no pool of helpful outside experts comparable to Open Source programmers. Rather than depend on outsiders, the military will often hire or train the best available experts in specialized military equipment (tanks or fighter planes) or applications (battlefield communications). Public disclosure of the defenses will then do little to improve the design of the defenses.

Under assumption (B3), the military will often be the organization affected directly by a vulnerability. There may be counter-measures for land mines (magnetic detectors) or for camouflaged machine guns (infrared detectors). If so, then the military generally has confidential channels for telling its own people what to do in response. There are few or no third parties who would benefit from disclosure of the vulnerability or know what to do about the vulnerability. (At least there are no third parties on "our side" that we want to tell.)

Turning briefly to the submarine situation during World War II, disclosure of the sailing time of a convoy helped attackers by revealing a vulnerability. Disclosure did little or nothing to help the Navy (the system designer for the defense) to protect the ships. Disclosure also did little to help other parties to defend themselves.[11] In this setting "loose lips" did indeed "sink ships"—the costs of disclosure outweighed the benefits.

Effects of assumptions B1, B2, and B3: in the military paradigm, the costs of disclosure of a vulnerability are high because attackers otherwise pay a high cost to learn of the vulnerability. The benefits of disclosure are low because outside designers are unlikely to improve the defenses and there are few or no third parties that the defenders wish to help through disclosure.

Taking the Open Source and military cases together, we can create a 2x2 matrix that visually shows the different effects of disclosure under the two paradigms. Under the Open Source assumptions, disclosure tends to improve the defense without helping the attackers. There are thus net benefits from disclosure. Under the military assumptions, the effects are reversed and there are net costs from disclosure.

[11] It is possible to imagine some assistance to third parties from disclosure. For instance, other ships might venture to sea if it becomes known that there is a convoy in another area that will draw the submarines' attacks. This benefit from disclosure, however, is likely to be outweighed by the harm to the convoy that becomes the target of the attack.

TABLE 1:
GREATER DISCLOSURE UP AND TO THE LEFT;
GREATER SECRECY DOWN AND TO THE RIGHT

		Help the Attackers Effect	
Help the Defenders Effect		*Low*	*High*
	High	*A: Open Source*	
	Low		*B: Military*

C. *Case C: The Information Sharing Paradigm*

The matrix also sheds light on when greater "information sharing" will improve security, such as the numerous information sharing provisions in the USA-PATRIOT Act[12] or proposals for the CIA and the FBI to share more of their data. Perhaps the easiest case to understand concerns sharing "watch lists" of suspected terrorists with defenders such as airport screeners, visa officers, and officials in other countries. Will greater disclosure of the watch list improve or harm security? The assumptions are:

(C1) Attackers may learn a lot from disclosure.

(C2) Disclosure may teach defenders how to design better systems.

(C3) Disclosure will allow more defenders to take protective actions.

[12] Uniting and Strengthening America By Providing Appropriate Tools Required to Intercept and Obstruct Terrorism (USA Patriot Act) Act of 2001, Pub. L. No. 107-56, 115 Stat. 272, Secs. 203, 326 [hereinafter USA Patriot Act]. *See also* Peter P. Swire, "Information Sharing, the Patriot Act, and Privacy," (presentation made Feb. 28, 2004), *at* www.peterswire.net.

The intuition for assumption (C1) is that broader dissemination of the watch list may tip off attackers who are on the list. The tip may occur either due to a mole (a rogue employee) or because the list is kept in an insecure place and gets leaked to the attackers. Persons who are on the list will then be on notice to avoid enforcement officials or to mask their identity. Persons who are not on the list will learn not to associate publicly with their colleagues on the watch list. Persons who are not on the list will also learn that they are "safe" and thus can fly on airplanes or otherwise get through screening processes. These "safe" people can then infiltrate defenses more effectively to spy or launch an attack.

The intuition for assumption (C2) is that broader use of watch lists, implemented by more defenders, may provide useful feedback for what sorts of watch lists are effective. A stronger intuition likely exists for assumption (C3). Putting the watch list into the hands of more defenders increases the likelihood of spotting and capturing the attacker. For instance, putting the picture of a "most wanted" criminal on television makes it harder for the criminal to escape. Especially where the criminal already knows that he or she is being chased, disclosure will help the defenders more than the criminal.

In practice, how the costs and benefits of disclosure compare will be an empirical question. Defenders will seek to create systems where defenders can effectively learn information while attackers cannot. As the number of defenders grows, however, it is less likely that every one of the defenders is trustworthy and every system containing the information is secure.[13] Information sharing is likely to have both costs and benefits, which will vary with the circumstances.

Effects of assumptions C1, C2, and C3: in the information sharing paradigm, there are significant costs and significant benefits from disclosure. The costs of disclosure may be high if attackers learn about the nature of the defense. The benefits of disclosure may be high if defenders can take additional, effective measures against the attackers.

D. Case D: The Public Domain

Another important possibility is that disclosure of a vulnerability will have low costs and low benefits. In some instances, a vulnerability is so minor that attackers will not be inclined to exploit it. More broadly,

[13] In some instances, technological measures may help get benefits from disclosure while minimizing the costs. The technological and institutional issues for doing so are beyond the scope of this paper. The most intense recent public debate has been about the CAPPS II system for screening airline passengers. *See, e.g.*, CENTER FOR DEMOCRACY AND TECHNOLOGY, TSA ISSUES SECOND PRIVACY ACT NOTICE EXPANDING AND NARROWING CAPPS II, (2003), *available at* http://www.cdt.org/headlines/20030731a.shtml.

in many settings the information is already in the public domain – the relevant information is already available to interested attackers and defenders. In such settings, the assumptions are:

(D1) Attackers will learn little or nothing from the disclosure.

(D2) System designers will learn little or nothing from the disclosure.

(D3) Other defenders may learn little or a significant amount from the disclosure.

An example of information in the public domain is the street map for Manhattan or Washington, D.C. Having a detailed and accurate street map is a great advantage for an attacker. In war-time, attackers crave good maps as they move into enemy territory. Good maps allow precise planning, facilitate coordinated attacks, and reduce the risk of hidden features that can booby-trap the assault. In response, defenders who know the terrain may remove street signs or take other measures to prevent the attackers from learning the area.

As part of the war on terrorism, it might thus be tempting for the United States to try to prevent terrorists from getting accurate street maps of potential targets such as Manhattan or Washington, D.C. The problem, however, is obvious. Detailed and accurate street maps of those cities are in the public domain, with innumerable copies in print and on the Internet. It would be very expensive even to try to hide the maps and such efforts would almost certainly be futile. In addition to these costs of trying to hide the maps, there would be substantial costs to all the legitimate users of the maps.

In terms of the three assumptions, assumption (D1) is that attackers would learn little or nothing new from a "disclosure" such as publishing an additional street map. Assumption (D2) is that the designers of the defense would learn little or nothing when a new street map is published. Assumption (D3) is that a new street map may in fact be of some use to other "defenders" such as legitimate users of the information including tourists, urban planners, and all others who rely on street maps.

From the other direction, efforts to hide or "re-classify" information will often be expensive and not very effective in an era of the Internet, on-line search engines, and archiving of information once it has been on the Internet.[14] The benefits of trying to hide the information will often be small because determined attackers will still have the information.

[14] For one informative discussion of the wealth of information available through the Google service, see Scott Granneman, *The Perils of Googling*, THE REGISTER, Mar. 10, 2004, *at* http://www.theregister.co.uk/content/55/36142.html.

The costs of trying to hide the information may be considerable, both in the effort to find and destroy copies that already exist and in the effect on legitimate users of the information.[15] Once a secret is exposed, it is often costly or impossible to put the genie back in the bottle.

Effects of assumptions (D1), (D2), and (D3): for information in the public domain, there are few or no costs from additional disclosure. There may be benefits from additional disclosure if additional legitimate users (defenders) learn from the disclosure. There are likely high costs from trying to hide data once it is in the public domain.

E. The 2x2 Matrix for When Disclosure Improves Security

With the addition of Case C on information sharing and Case D on the public domain, each cell of the 2x2 matrix has been filled in. Table 2 shows the result:

Table 2:
Greater Disclosure Up and to the Left;
Greater Secrecy Down and to the Right

		Help the Attackers Effect	
Help the Defenders Effect		*Low*	*High*
	High	*A: Open Source*	*C: Information Sharing*
	Low	*D: Public Domain*	*B: Military*

[15] The discussion here focuses only on the extent to which the disclosure will help or hinder the attackers. Efforts to censor information in the public domain also can obviously raise serious First Amendment and other problems. Eugene Volokh has written an excellent analysis of these issues, in an approach that is congruent in a number of respects with the analysis in this paper. Eugene Volokh, *Crime-Facilitating Speech*, (2004) (unpublished manuscript, on file with author).

At this stage, a few comments will help to emphasize what is and is not accomplished by Table 2. First, a chief goal of the table is to organize our current thinking about the dueling approaches of disclosure ("no security through obscurity") and secrecy ("loose lips sink ships"). By clarifying the assumptions underlying those two scenarios, the table also reveals the assumptions underlying two other common scenarios – Information Sharing and the Public Domain. Second, the table simplifies reality by showing a binary split between high and low effects of helping the attackers and improving the defense. In reality, there is a continuum between high and low effects. Real-world examples will range along the two dimensions. Third, the table is based on *assumptions* about the effects of disclosure on attackers and defenders. Conclusions about the desirability about a disclosure will depend on how valid the assumptions are in a given setting.

II. THE KEY REASONS COMPUTER AND NETWORK SECURITY MAY VARY FROM OTHER SECURITY PROBLEMS

In the legal academy, there has been a lively debate about the extent to which cyberspace (and the law of cyberspace) is different from the physical world (and the law of the physical world). For instance, writers such as David Post and David Johnson have stressed the uniqueness of the Internet, while writers such as Frank Easterbrook and Jack Goldsmith have stressed how the law of the Internet is fundamentally similar to previous legal issues.[16] The topic of this section is to examine the extent and nature of the differences between computer and network security, on the one hand, and the military and other traditional security problems of the physical world, on the other.

The conclusion here is that there is no *logical* or *necessary* difference between cybersecurity and physical security. One can generate examples where the nature of the security challenge and the optimal degree of disclosure are the same regardless of what is being protected. Nonetheless, the claim here is that there are reasons why there are *commonly* important differences between cybersecurity and physical security. These differences, I believe, contribute a great deal to why so many cybersecurity experts intuitively believe in "no security through obscurity" while so many military and other physical security experts intuitively believe that "loose lips sink ships."

[16] *See* Jack L. Goldsmith, *Against Cyberanarchy*, 65 U. CHI. L. REV. 1199, 1199 n.3 (1998) (collecting citations to works of Post, Johnson, and others who stress uniqueness of cyberspace law). *But see* Frank H. Easterbrook, *Cyberspace and the Law of the Horse*, 1996 U. CHI. LEGAL. F. 207 (1996).

A. Hiddenness and the First-Time Attack

Here is an organizing concept for when hiddenness helps security: a hidden feature is more likely to be effective against the first attack, but less likely to be effective against repeated attacks. Consider an example from the physical world. A fort is protected by a simple security device that relies on hiddenness. On the path up to the fort there is a pit covered by leaves, with a sharpened stick at the bottom. The first time an attacker comes up the path, the attacker might fall into the pit. Even if hiddenness works against the first attacker, however, later attackers will likely not "fall" for the same trick. The later attackers may know where the pit is, or they may come equipped with sticks that probe the path so that they don't fall in. In this simple example, using obscurity may work against the first attacker, but is unlikely to work once the attackers learn to watch for the hidden pit.

The concept of the first-time attack can be generalized. Consider a "hidden" defensive feature as one that is not known initially to any attacker. The effectiveness of hiddenness will be a function of five variables:

> (1) The effectiveness of the defensive feature at stopping the first attack. ("*E"* for effectiveness.) (2) The number of attacks. ("*N*" for number of attacks.) (3) The extent to which an attacker learns from previous attacks. ("*L*" for the learning that occurs.) (4) The extent to which the attacker communicates this learning to other attackers. ("*C*" for communication.) (5) The extent to which the defenders can effectively alter the defensive feature before the next attack. ("*A*" for alteration of the defense.) Note that the alteration may come from the system designer/defender (A-D). The proposed alteration may also come from third parties who learn how to fix the vulnerability (A-T), such as when an Open Source programmer designs a patch.

The effectiveness of hiddenness will vary directly with greater initial effectiveness (*E*) and greater ability by the designer to alter the defense (*A-D*). It will vary inversely with the number of attacks (*N*), the degree of learning by attackers (*L*), the ability of attackers to communicate (*C*), and the ability of third parties to alter the defense (*A-T*). When the effects of *N*, *L*, and *C* grow very large, there will be no usefulness of hiding the defensive feature. All attackers will then know everything about the "hidden" feature.[17]

[17] The discussion here does not present a detailed mathematical model of how hiddenness contributes to security. Identification of the five variables, however, should enable those who are mathematically more skilled than I am to build such a model. As suggested in conversation by Rena Mears, the approach here implicitly assumes a calculus function where the effectiveness of hiddenness goes to zero as the number of attacks approaches infinity

The military and Open Source examples explained earlier illustrate how the effectiveness of hiddenness can vary depending on the five variables. Start with the example of camouflaged machine guns guarding a hill, where the effect of hiddenness is the difference between announcing the location of the machine guns and keeping them hidden. Initially, the attackers do not know where the machine guns are hidden. *E*, the effectiveness of the defense, will likely be high against infantry attackers because the hidden guns will make it hard for attackers to find a safe path up the hill.[18] *N*, the number of attacks, will be low. Each attack is a major event that is costly in terms of casualties. *L*, or learning, will vary depending on the ability of an individual attacker to get back safely from the first attack or go around the machine gun nest. If all the attackers are killed in the attack, then *L* will be zero. *C*, or the ability to communicate, will vary depending on the ability of any individual attacker to tell the rest of the troops about the location of the hidden guns. If the attackers have radios, then there will be a high *C* because they can tell their comrades what locations to avoid. If the attackers have to rely on word-of-mouth, then *C* will be low and the defense may have time to set up a new ambush in time for the second attack.

Pulling these observations together, each attack on the hidden machine guns is very expensive for the attackers. Hiddenness benefits the defender in the first attack. The number of attacks will be small. (would there be even three or four charges against a well-defended hill?) Attackers may not learn quickly about the hidden defenses, may find it difficult to communicate their learning to the other attackers, and may face a changed defense by the time they launch their next attack. For all of these reasons, hiddenness will benefit the defense.

Under the assumptions used thus far for Open Source software, hiddenness will be much less effective. It is possible that the initial effectiveness of a defensive trick, *E*, will be substantial. The number of attacks, *N*, will quite possibly be high. Malicious hackers can probe for weaknesses in a software product over and over again. The attackers learn (*L*) from the attacks, such as by seeing whether they can gain control over the software. Attackers can communicate (*C*) about flaws, such as by posting their exploits to web sites to let other attackers know about the flaws.

(assuming a positive value for *L* and *C*, and also assuming the effect of *L* and *C* in helping attackers outweighs the effect of alterations in helping defenders). *See* Conversation with Rena Mears, Partner, Deloitte & Touche (Feb. 20, 2004).

[18] The example here assumes foot soldiers charging up a hill against machine guns. If the attack is made by heavy tanks, then ordinary machine guns will not stop the attack. For the tank attack, the value of *E*, the initial effectiveness, would be low.

Under these assumptions, each attack on a software program is very cheap – attackers can probe the program over and over again from the comfort of their own homes or computer labs. They learn about flaws and tell others about flaws. Very quickly, under these assumptions, the hidden defense is exposed to the world. Thus, there is "no security through obscurity."

The possibility of altering the defense also works differently than for the physical attack against machine guns. In the machine gun setting, the defense may be able to move the guns between each attack. If that is true, then the second ambush may be as effective as the first, and hiddenness once more favors the defender. Under the Open Source assumptions, disclosure of the vulnerability actually increases A, the likelihood of effective alteration of the defense. The idea is that other Open Source programmers will come forward to write a patch for the vulnerability. In terms of hiddenness, improved protection against the next attack works in opposite ways for the machine gun and Open Source examples.

B. Uniqueness of the Defense

How should we refer to the effect of the five variables? Using the term "first-time attack" has the advantage of communicating to a wide audience. Through understanding ordinary English, a reader can grasp the idea that a hidden trick may work against the first attack but fail against the 1000th attack. The problem with the term "first-time attack," however, is generalizing the effect to "second-time attacks" (hiddenness may still work very well), "twentieth-time attacks" (hard to know how well hiddenness will work), and "nth-time attacks" (the hidden features will quite possibly be discovered).

This article will use the word "uniqueness" to refer to the usefulness of hiddenness for the defense. Despite the possible complaints of English teachers,[19] this article discusses uniqueness as a function, varying from "unique" or "entirely unique" down through "somewhat unique" to "not unique at all."[20] The function for uniqueness (*U*), or the usefulness of hiddenness for the defense, is thus:

$$U = \mathrm{f}\,(E, N, L, C, A)$$

Under the terminology employed here, "high uniqueness" refers to situations where hiddenness is effective, due to a combination of high values of initial effectiveness (*E*) and ability to alter the defense (*A*) and low values for the number of attacks (*N*), learning from previous attacks (*L*), and communication among attackers (*C*). "Low uniqueness" refers to situations where the values are reversed.

C. *Why Low Uniqueness May Be Common for Computer and Network Security*

Important areas of computer and network security include: perimeter defense such as firewalls; mass-market software, including video games; and encryption. For each of these areas there will often be a low degree of uniqueness, so secrecy is unlikely to be very effective.

1. Firewalls

One meaning of "no security through obscurity" on the Internet is that it is a bad strategy to try to hide: a new system is likely to be disvoered and probed almost as soon as it comes on line. More generally there is a plausible case that firewalls are subject to a large number of attacks (*N*), considerable learning by attackers (*L*), and effective communications among attackers (*C*). Using the Internet, attackers can probe a firewall from anywhere on the planet. They can attack again and again at low cost, trying various combinations of attacks until they find one that works. They can then tell other attackers about the

[19] One web page lists "errors" in English usage, and says: "'Unique' singles out one of a kind. That 'un' at the beginning is a form of 'one.' A thing is unique (the only one of its kind) or it is not. Something may be almost unique (there are very few like it), but nothing is 'very unique.'" http://www.wsu.edu:8080/~brians/errors/unique.html (last visited July 17, 2004).

[20] When I presented this paper at a conference at the Stanford Law School, Bruce Schneier and Matt Blaze both suggested the term "instance" to refer to what I am here calling "uniqueness." I have chosen the latter term for two main reasons. First, "instance" has so many uses in English that it may be confusing to readers for it to have a more technical definition. Second, my sense is that readers will intuitively understand the idea of different degrees of uniqueness.

vulnerability, such as by posting a script of the attack to a web site or e-mail list. Even unskilled "script kiddies"[21] may then be able to use the attack to pierce that firewall or other firewalls that use the same defenses.

Comparison with an attack on a walled city illuminates the way that computer and physical attacks are both similar and different. The similarities between a computer firewall and a medieval city wall are easy to see. A strong barrier is designed to allow friends to enter but keep foes out. Either type of defense can be set to various levels of security. In times of peace, a city gate may allow anyone to enter, with guards on hand to handle anyone suspicious. At a higher level of alert, guards might check the credentials of each person before entering the city. During a siege, the gates might be closed completely, barring all entry. Additional security might exist within the city wall. For instance, the armory (containing weapons), the mint (containing treasure), and the castle keep (containing the ruler) all would have additional protections against entry.

A company's firewall is similar. For non-essential systems most messages will be allowed entry. For secure systems, a password or other credential is required. Under severe conditions, such as a distributed denial of service attack, all messages may be blocked from entering the company's system. Additional security will exist for priority functions, such as the system security (the armory), the corporate treasury (the mint), and the root directory (the ruler's residence).

Along with these similarities, it is logically possible for attacks against a physical wall to have high *N*, *L*, and *C*. For a long and badly defended wall, for instance, intruders might repeatedly probe for weak spots, learn about vulnerabilities, and tell fellow attackers where to enter.[22]

Many attacks against a city wall, however, do not fit that pattern. In medieval warfare, an attack against a walled city was a major event in which many people might die. Any hidden trick by the defenders might cost attackers' lives or save defenders' lives before the attackers learned how to counter the trick. The number of attacks was low, attackers might not survive to tell about weak spots, and communication back to the attacking generals was rudimentary. Similarly, any hidden

[21] "Script kiddies" are unskilled programmers who merely follow a script rather than understanding how to write code themselves. *See, e.g.*, THE JARGON DICTIONARY, SCRIPT KIDDIES, *at* http://info.astrian.net/jargon/terms/s/script_kiddies.html (last visited July 17, 2004) (defining script kiddies as "the lowest form of cracker; script kiddies do mischief with scripts and programs written by others, often without understanding the exploit.").

[22] An example of a physical barrier with high *N*, *L*, and *C* might be the United States border with Mexico. There are many persons who seek to cross the border, there are professionals who learn the soft spots in the defenses, and others who wish to cross the border learn from earlier successes.

weaknesses might not be revealed in time to help the attack. In short, *N*, *L*, and *C* would all be low.

In sum, low levels of N, L, and C likely meant that medieval city walls had high uniqueness – secrecy was likely to be a useful tool. Firewalls using standard software likely have low uniqueness due to the high levels of N, L, and C.[23]

2. Mass-market software and computer games

Another major topic of modern computer security is how to protect standardized software against hackers. Popular products may be on thousands or millions of desktops. Designers of standardized software might try to use hiddenness to stop the hackers. For instance, the designer might have a program freeze up permanently if a user hacked into inappropriate portions of the software. This kind of defense would be similar to falling into the pit covered with leaves – the attacker who goes into the wrong place never comes out again.

This hiddenness will often not work well, however, for mass-market software. Suppose, for instance, that there are a dozen paths for hacking a piece of code to do something forbidden such as send a virus or make illegal copies. Suppose the designer puts traps on eleven of the twelve, to freeze up the program permanently if a hacker trespasses into the wrong part of the code. Suppose further that the designer leaves the twelfth path free so that the designer can get back in to rewrite the code.

This sort of defense would work reasonably well against a one-time attack. In the physical world, an attacker would face a grave risk (11 out of 12 attempts) of falling into the pit and getting injured. Similarly, in the computer world, a hacker who can get only one copy of the program, and who needs that program to keep functioning, could find it too risky to fool around with the program and likely have it freeze into uselessness. In practice, though, a hacker can often find ways to create a backup copy or find other ways to test the software repeatedly. This hacker can systematically try one possible attack after another until something works – a high *N* and *L*. Meanwhile, other hackers around the world also try their favorite attacks, and the hackers can communicate amongst themselves when they find a vulnerability – a high *C*.

[23] Despite the intuition that firewalls have low uniqueness, I have talked with some computer security experts who build higher uniqueness into their own firewalls. Even for some experts who support the idea of "no security through obscurity" there is an understanding that putting some hidden tricks into a defensive system such as a firewall can be helpful. Notably, the hidden or subtle changes can stop attacks by "script kiddies" and others who are not able to modify their attacks in the face of a new defense.

The combination of high *N*, *L*, and *C* also exist for computer and video games today when players try to "beat the game."[24] "Beating the game" is a (presumably) innocent version of hacking a software system – users ultimately reach their goal of gaining control over the software. An old-fashioned (although perhaps satisfying) way to "beat the game" is to keep trying by yourself until you overcome all the obstacles. As an alternative, video game players today can also enlist a global network of fellow aficionados. Web sites appear almost instantly after release of a game. The sites offer "secrets" (press the third brick on the left to get a magic sword), "walk throughs" (on Level 13 here are the seven things you have to do before you attack the dragon), and even "cheats" (if you enter this code, your player will become invulnerable to all attacks and as strong as Superman). Translated back into the language of computer security, there is a high number of attacks, *N* – just ask the parents. Users learn from experience and communicate that learning – a high *L* and *C*. A hidden measure by the game designers will not stay hidden for long.

In summary, where there are high levels of N, L, and C for attacks on mass-market software, there will tend to be low uniqueness and little "security through obscurity."

3. Encryption

Encryption is a third major area of modern computer security, along with system defense (firewalls) and defending software. The word "encryption" comes from the Greek word for "hidden," so it might seem exceedingly odd to say that being hidden does not work well for encryption.[25] Yet, in the sense used in this article, that is precisely the claim. The question, for our purposes, is whether hiddenness paired with encryption that suffers from vulnerabilities will succeed, or whether instead security can be provided only by strong encryption, i.e., encryption that is successful even when the attacker knows the method used to encrypt the message.

[24] This paragraph is based on insights from my sons Nathan and Jesse Swire, now 15 and 13.

[25] For excellent historical introductions to encryption, see DAVID KAHN, THE CODEBREAKERS: THE STORY OF SECRET WRITING (1996); *See also* SIMON SINGH, THE CODE BOOK: THE EVOLUTION OF SECRECY FROM MARY QUEEN OF SCOTS TO QUANTUM CRYPTOGRAPHY (1999).

Modern cryptographers are likely the most avid believers that there is no security through obscurity. Cryptographic authority Bruce Schneier has stated:

> A basic rule of cryptography is to use published, public, algorithms and protocols. This principle was first stated in 1883 by Auguste Kerckhoffs: in a well-designed cryptographic system, only the key needs to be secret; there should be no secrecy in the algorithm. Modern cryptographers have embraced this principle, calling anything else "security by obscurity." Any system that tries to keep its algorithms secret for security reasons is quickly dismissed by the community, and referred to as "snake oil" or even worse.[26]

Schneier, with his discussion of "snake oil," highlights the risk that a vendor will dupe purchasers of an allegedly secure system. Once the system is exposed to attack, however, the system may have only weak protections, and all of the communications of the purchaser may thus be exposed to view. Having "published, public, algorithms and protocols" is thus an important consumer protection against the vendor who tries to hide the vulnerabilities of a weak system.

A second reason for the cryptographers' belief in openness is that a secret is unlikely to remain secret when known to a large number of people. Cryptography today is used by an enormous number of users on the Internet. In earlier times, by contrast, encryption was used by far fewer persons, most prominently by diplomats and the military. Encryption became more widespread when people wished to send a lot of important messages through a channel where other people could see or hear the message. In times when the post was not secure, letter writers used encryption. In the days of the telegraph, many businesses used encryption to keep their commercial secrets away from the eyes of the telegraph operators. For radio communications, anyone with a receiver could hear the message. Most famously, German submarines in World War II used the Enigma system when radioing back to headquarters. Allied cryptographers learned to break the system after enormous effort, helping to win the war and more or less inventing the computer as a by-product.

The need for encryption is thus not new with the Internet. But the Internet has been accompanied by an enormous increase in the need for and use of encryption by ordinary people and businesses. The Internet is

[26] Bruce Schneier, *Secrecy, Security, and Obscurity*, CRYPTOGRAM NEWSL. (May 15, 2002) *at* http://www.schneier.com/crypto-gram-0205.html. Schneier returned to these issues in BEYOND FEAR: THINKING SENSIBLY ABOUT SECURITY IN AN UNCERTAIN WORLD 126-32 (2003).

a famously "open" system.[27] A message from Alice to Bob is typically routed through many computers on its way through the Internet. A hacker might be in control of any of those computers. The hacker might make a copy of all the messages coming through the system and then comb through the messages looking for any that have commercial, diplomatic, or military value. In response, Alice and Bob need to encrypt their important messages, containing credit card numbers, trade secrets, large transfers of currency, and anything else they don't want the hacker to read and copy.

The Internet does more than increase the number of messages that use encryption. The Internet has also accelerated demand for public-key encryption approaches that permit anyone to send an encrypted message to anyone else. The basic idea of a public-key system is that a user, Alice, can send a message to a recipient, Bob, whom she has never met before.[28] She uses Bob's public key to encrypt her message. Bob can then decrypt it using his private key. The public key can be posted on the Internet or otherwise revealed to the world. The private key is kept secret by Bob and not made known to attackers. The combination of many messages through insecure channels (the Internet) and many users who wish to communicate securely with each other (as in E-commerce) has meant that an unprecedented number of individuals rely on cryptosystems[29] that are widely deployed.

[27] *See generally* Jane Kaufman Winn, O*pen Systems, Free Markets, and Regulation of Internet Commerce*, 72 TUL. L. REV. 1177 (1998) (discussing the openness of the Internet).

[28] For further discussion, see Bruce Schneier, APPLIED CRYPTOGRAPHY ch. 19 (2d ed. 1996).

[29] Modern encryption draws a distinction between the "cryptosystem" and the "key." The cryptosystem is a mathematical technique that has a standard way to re-arrange symbols ("put every second letter in front of the letter before it") and substitute one symbol for another ("change each letter A into the number 1"). The most widely-used modern cryptosystems publish the algorithm for converting between plaintext (readable English) and ciphertext (the message as transmitted in its encrypted form). A well-known example is the RSA algorithm developed in 1978 by mathematicians Ronald Rivest, Avi Shamir, and Leonard Adleman. The security of the RSA cryptosystem depends on a mathematical algorithm that is easy to calculate in one direction (when one encrypts the message) but extremely difficult to calculate in the other direction (when an unauthorized person tries to decrypt the message.) For the mathematical basis of the RSA algorithm, created in 1978, see *What is the RSA Cryptosystem? at* http://www.rsasecurity.com/rsalabs/node.asp?id=2214 (last visited Aug. 5, 2004).

The security of the RSA cryptosystem also depends on each user having a secret key to turn ciphertext back into plaintext. The idea of a key is simple enough. Suppose that the cryptosystem turns each letter into a number, such as A=1, B=2, C=3, and so on. There are 26 possible starting points, such as A=25, B=26, C=1, and so on. In this simplified example, the cryptosystem is a regular pattern for turning letters into numbers. The key is knowing how to begin the calculation, by knowing which number corresponds to the letter A. In actual cryptosystems, the key is a long chain of randomized numbers. Attackers who do not have the key then need to try every possible combination of numbers until a key fits the lock (decrypts this plaintext). Trying each of the combinations, which can easily number in the billions, Given this understanding of today's networked encryption, we can now better understand why modern cryptographers believe there is no security through obscurity. Because so many communications flow through the Internet and can be read by hackers, the number of attacks, *N*, is extremely high. If *L* and *C* are even slightly positive, then attackers will learn about the vulnerabilities in a method for encrypting messages and communicate about those vulnerabilities to others. The response by cryptographers is to use methods for encryption that do not rely on secrecy. Instead, cryptographers increase the length of the secret key to try to make brute force attacks prohibitively costly.

The combined effects of *N*, *L*, and *C* mean that the cost of disclosure of the cryptosystem – the help-the-attackers effect – is low. The benefit of disclosure to defenders is also likely to be high. For one thing, use of a public-key algorithm means that myriad users can easily send encrypted messages to each other. In addition, there is likely a high value for *A*, the ability of defenders to improve the defensive system. The rise of the Internet and the spread of public-key encryption has led the number of encryption experts to grow rapidly in recent years. The likelihood of improved defenses is thus substantial: "The long history of cryptography and cryptanalysis has shown time and time again that open discussion and analysis of algorithms exposes weaknesses not thought of by the original authors, and thereby leads to better and more secure algorithms."[30]

Before leaving the topic of encryption, it might be useful to see how this conclusion – the advantage of an open cryptosystem – would have been less true in Roman or Medieval times. In that setting, there likely would have been lower *N*, *L*, *C*, and *A*. The number of encrypted messages subject to interception would have been far lower than on the Internet. The sophistication of those intercepting the messages would have been lower. Slow communications would have meant that other attackers would have learned very slowly, if at all, from the breakthrough by one attacker. In addition, the chances of "outside cryptographic experts" improving the system would have been low. All of these variables would therefore have pointed toward the usefulness of a hidden cryptosystem, in contrast to conditions today.[31]

trillions, and up, is called a "brute force attack." An attacker who can try every single possible key will eventually be able to read the code. The response by those who build cryptosystems is to try to make the number of combinations so large that no available computer can try all the combinations.

[30] BUSH & BELLOVIN, *supra* note 9.

[31] In comments on an earlier draft, cryptographer Susan Landau disagreed with the discussion of the role of hiddenness in earlier times. She mentioned a 14th Century Arabic encyclopedia, the Subh al-a 'sha, that contained sophisticated mathematical techniques for breaking ciphers. In response, the claim here is that secrecy is more likely to have net benefits

In summary, secrecy in modern cryptosystems is unlikely to be useful due to high N, L, C, and A. Modern encryption relies, however, on strict secrecy for private keys.

III. Relaxing the Open Source Assumptions – Computer and Network Security in the Real World

Section II sought to explain why computer security experts so often believe there is no security through obscurity. Firewalls, mass-market software, and encryption are major topics for computer and network security. In each setting, there are typically high values for number of attacks (*N*), learning by attackers (*L*), and communication among attackers (*C*). Secrecy is of relatively little use in settings with high *N*, *L*, and *C* – attackers quickly learn about the hidden tricks. By contrast, many physical-world security settings have lower values for *N*, *L*, and *C*. In these settings of persistent and higher uniqueness, secrecy is of greater value to the defense.

Section II thus solidified the assumptions of the Open Source paradigm, that (1) disclosure will offer little or no help to attackers; (2) disclosure will tend to upgrade the design of defenses; and (3) disclosure will spread effective defenses to third parties. High levels of *N*, *L*, and *C* strengthen the first assumption, because attackers will quickly learn about secrets. Alterations (*A*) from outside experts, in cryptosystems and elsewhere, fit with the second assumption. Finally, high levels of *A* and *C* will alert other defenders to vulnerabilities under the third assumption.

After this reinforcement of the Open Source assumptions, Section III will now try to test the assumptions in the real world. In practice, secrecy will often be of greater use than suggested by the assumptions of the Open Source paradigm.

A. *The Assumption that Disclosure Will Not Help the Attackers*

The first assumption in the Open Source paradigm is that disclosure will provide little or no help to the attackers. The assumption is that there are many capable persons who are willing and able to launch attacks against firewalls, mass-market software, and cryptosystems.

in situations with lower *N*, *L*, *C*, and *A*. Where attackers such as users of that encyclopedia have sophisticated techniques, they will have higher *L*, reducing the effectiveness of secrecy. The claim in the text is that earlier periods generally had far lower *N*, *L*, and *C* than would attacks today on a widely-used cryptosystem on the Internet. Modern attackers will thus be more efficient at overcoming hidden defenses (due to today's higher learning) and modern defenders will be more likely to get suggestions for useful alterations (due to today's larger group of potentially helpful cryptographic experts). There will thus be higher expected benefits today of disclosure of the cryptosystem.

To scrutinize this assumption, it is important first to develop the intuition that the public domain of information is expanding in a world of search engines such as Google. Next, disclosure can sometimes help defenders when the disclosure deters attacks. Third, the case for disclosure of private keys, such as cryptographic keys, is especially weak. Fourth, the area of surveillance is subject to a different analysis. Finally, the discussion turns to a more specific discussion of the extent to which attackers already know about how to launch effective attacks against firewalls, mass-market software, and cryptosystems.

1. The Enlargement of the Public Domain in a World of Search Engines

Do attackers know specific facts about defenders? The answer today, in a world of the Internet and search engines, is that the cost of doing searches has gone way down. Many facts that were impossible or costly to find in the past are easy to find today.

All readers of this article know this to some extent, but it is helpful to flesh out some of the reasons that so much more information is today in the public domain. The Internet itself has only recently expanded beyond the domain of DARPA[32] and the academic community. Indeed, it was not until 1992 that the terms of service for the Internet changed to permit commercial activity on the Internet.[33] The growth in commercial activity coincided with the incredible expansion of Internet usage, so that ordinary people all over the world could find out information, at no cost, about a huge range of topics. Search engines have made it trivially easy to search through the many web sites to find specific information. Google was launched in 1998 and indexed 30 million items at that time. Today, it indexes over 6 billion items.[34]

At a simple yet powerful level, the ubiquity of search engines (and the other research tools of the Information Age) increases the knowledge available to attackers. Attackers can correlate information from diverse

[32] DARPA is the Defense Advanced Research Projects Agency, which played a key role in fostering the early stages of the Internet. *See* Michael Hauben, *History of ARPANET; Behind the Net - The untold history of the ARPANET; Or - The "Open" History of the ARPANET/Internet*, *at* http://www.dei.isep.ipp.pt/docs/arpa.html (last visited July 17, 2004).

[33] The Scientific and Advanced Technology Act of 1992, signed into law on October 23, 1992, "subtly modified [the National Science Foundation's] authority to support computer networks that are not limited to research and education." NAT'L SCI. FOUND., OFFICE OF INSPECTOR GENERAL, REVIEW OF NSFNET, (Mar. 23, 1993) (citing 42 U.S.C. § 1862(g)). This change was one important legal step toward development of commercial activity over what is now called the Internet.

[34] Robert Weisman, *Investors Monitoring Climate for Google IPO*, MIAMI-HERALD.COM, (Mar. 21, 2004) *at* http://www.miami.com/mld/miamiherald/business/national/8243019.htm.

sources to infer facts that are themselves not explicitly made public. Attackers can communicate with other attackers through blogs,[35] web sites, and global, free e-mail. Search engines are extremely useful. For instance, you can pick the mass-market software or firewall you wish to attack and search on Google for the name of the product and "bugs" or "vulnerabilities." If you do so, you may find a patch that has already been announced for the known bug. In launching actual attacks, you are also likely to discover that a large portion of the product's users have not installed the patch.

The increase of information in the public domain increases the set of instances where the Open Source paradigm is a better approximation than the military paradigm. More often than before, disclosure of a security flaw will add little or nothing to attackers' knowledge. It will be harder to keep many things secret, because attackers will be able to infer the truth from other available information. At the very least, the ubiquity of search engines increases the costs of trying to keep information out of the hands of attackers.[36]

In summary, the growth of the Internet and of search engines means that the optimal solution often shifts toward openness in weighing the costs and benefits of disclosure. In many instances, the help-the-attacker effect is likely to be low, while the costs to defenders of trying to keep secrets will have risen.

[35] For an early discussion of the legal implications of weblogs, or "blogs," see Attiya Malik, *Are You Content with the Content? Intellectual Property Implications of Weblog Publishing*, 21 J. MARSHALL J. COMPUTER & INFO. L. 439 (2003).

[36] There is a growing literature that laments the shrinking of the public domain. *See, e.g.*, Lawrence Lessig, THE FUTURE OF IDEAS: THE FATE OF THE COMMONS IN A CONNECTED WORLD (2002); James Boyle, *The Second Enclosure Movement and the Construction of the Public Domain*, 66 L. & CONTEMP. PROBS. 33 (2003). This literature emphasizes, for instance, the ways in which copyright and other intellectual property rules have expanded. Even though copyright law itself does not apply to facts, some actual and proposed legal developments could reduce the set of facts available to the public. For instance, the anti-circumvention provisions of the Digital Millennium Copyright Act, 42 U.S.C. § 1201, can make it illegal to access facts that are in a format protected by anti-circumvention measures. In addition, there have been repeated legislative attempts in the United States to enact *sui generis* database protection, which would create new limits on the ability of users to reproduce facts in certain databases. Jonathan Band, *New Theories of Database Protection*, MANAGING INTELL. PROP. (Mar. 2003), *available at* http://www.legalmediagroup.com/mip/default.asp?Page=1&SID=1835.

Notwithstanding these concerns, the argument here is that the development of the Internet and of search engines has made available an increased range of factual information at lower cost than previously. Especially in the wake of the attacks of September 11, there have been some measures by the U.S. government to reduce the information available to the public. Edward Lee, *The Public's Domain: The Evolution of Legal Restraints on the Government's Power to Control Public Access Through Secrecy or Intellectual Property*, 55 HASTINGS L.J. 91 (2003). Despite these changes, vastly more security information is available today to a teenage hacker or a foreign terrorist than would have been true before the rise of the Internet.

2. Deterrence as a Result of Disclosure

Up until this point, the focus has been on how disclosure may reveal vulnerabilities and thus help the attackers. More generally, the analysis should include the full range of ways that attackers might respond to disclosure about the defense. One principal way that disclosure can help the defense is through deterrence – the effect of disclosure on *reducing* the likelihood of attack.

The distinction between "strong" and "weak" is likely the main axis for when disclosure will create deterrence. Attackers who see a "strong" defense will tend to be less likely to attack. Attackers who see a "weak" defense are more likely to believe that they will be able to overcome the defenses. This effect is not always true – a "strong" defense, for instance, might be a clue to an attacker that something valuable is contained inside.[37] Nonetheless, the appearance of a "strong" defense is generally a good predictor for the magnitude of deterrence.[38]

Deterrence can exist because the defense is strong in an absolute sense. In such circumstances, the defender will perceive that the costs of the attack are greater than the benefits. For example, assume that in the physical world there is a high fence, topped with razor wire and with surveillance cameras in clear sight. A potential trespasser who sees this defense may estimate that it will be difficult to climb the fence, dangerous to get over the razor wire, and risky in terms of being detected and caught.

Deterrence can also exist in a relative sense. There is an old story about two hikers in the woods who see a dangerous bear rushing toward them. One of the hikers turns around and starts running. The other hiker asks why he is running when everyone knows that bears can run faster than people. The first hiker responds: "I don't have to run faster than the bear. I just have to run faster than you." In terms of deterrence, a house with bars on the windows and large locks on the front door may simply be more trouble to attack than a neighboring house that lacks these features. The visible defense measures, in such circumstances, may shift the risk of attack to the neighboring house.

[37] As another example where deterrence would not succeed, some attackers might be attracted to a strongly defended target simply because it is strongly defended. Just as medieval knights sought glory by attacking famous champions, modern-day hackers sometimes seek "hacker glory" by attacking systems that are thought to be highly secure.

[38] The "strong"/"weak" distinction was first suggested to me by Jim Steinberg, who served as Deputy National Security Advisor under President Clinton. The fact that the suggestion came from a person steeped in national security issues suggests that the deterrence effect may implicitly be an important way that military and national security experts decide when disclosure will help security.

An essential element to successful deterrence is that the attackers know about the strong defense. This element was memorably missing in the movie *Dr. Strangelove*, where the Soviet Union failed to tell the rest of the world about the existence of a doomsday device that would be triggered by any nuclear attack.[39] When one nuclear bomb was accidentally used, the entire world was destroyed. The complete failure of communication in that instance drives home the point – it is the perception of the defense by the attackers that is key to deterrence.

In summary, the effects of disclosure on security include the deterrent effect on attacks (a help-the-defense effect) as well as the help-the-attackers effect discussed previously. The chief predictor of deterrence is the extent to which attackers perceive the defense as strong.

3. Don't Disclose Private Keys, Passwords, or Combinations to a Safe

The discussion of encryption, above, drew a sharp distinction between the cryptosystem and the private key. Modern cryptographers generally support "no security through obscurity" and favor disclosure of the cryptosystem. They also support secrecy for the private key or password. Modern cryptographic systems feature a high initial effectiveness for the cryptosystem (*E*). They also are resistant to a high number of attacks (*N*). The private keys are long enough to require brute force attacks that are too lengthy for attackers to undertake.

A similar analysis applies to physical protections such as a combination safe. The initial effectiveness (*E*) is high because attackers cannot easily get through the metal skin of the safe. Next, the combination of the safe is complicated enough to make a brute force attack difficult (resistant to a high *N*). A complex combination can be very effective—bank robbers typically do not wish to stay in the bank vault long enough to try every possible combination to open the safe.

For passwords, a good practice is to make it difficult for attackers to guess the password. Programs to guess passwords are easily available on the Internet.[40] In response, good practice is to require a password to include symbols and numbers in addition to letters. That practice increases the initial effectiveness (*E*) by forcing users not to use the defaults that come with software or common terms such as "password." Use of different characters in the password increases the number of

[39] DR. STRANGELOVE OR: HOW I LEARNED TO STOP WORRYING AND LOVE THE BOMB (Columbia Tri-Star 1964).

[40] For the person interested in testing this, simply use a search engine with terms such as "password hacker."

attacks (*N*) needed to guess the password. In addition, altering the password periodically (*A*) reduces the likelihood that attackers can continue to take advantage of one successful attack.

For all three defenses – the private key, the combination to the safe, and the password – there is a large help-the-attacker effect from disclosure of the secret information. All three defenses are designed to frustrate a brute force attack by having too many possible combinations. If there is disclosure of the secret key, then the entire defensive strategy falls apart.

Is there a help-the-defender effect from disclosure? Almost always, the answer is no. In these examples, the defenders are relying on fundamentally sound defenses, such as a strong cryptosystem or a heavy metal safe. These defenses will not be designed better just because one user's key or one safe's combination is revealed.

In summary, there is a large help-the-attacker effect from disclosure of a private key, combination to a safe, or password. There is usually no help-the-defender effect. Even for supporters of "no security through obscurity," this sort of information should stay secret.[41]

4. Why Secret Surveillance May Improve Security

The next question is whether it improves security to reveal surveillance techniques used by defenders. Under the Open Source paradigm, one might believe that disclosure will help the defenders because outside experts will suggest improvements to the surveillance system. In addition, the Open Source paradigm would suggest that attackers already know or will readily learn about the defenses of a system, so that disclosure will not help the attackers. The intuitions of intelligence experts are precisely the opposite. These experts believe that it is imperative to keep secret the sources and methods used for intelligence gathering.

The model for uniqueness shows why the latter view is usually better for achieving security. The key factual point is that attackers usually learn little or nothing about surveillance (low *L*) from their attacks. As the level of *L* approaches zero, then attackers do not learn

[41] One can imagine a couple of settings where disclosure of the private key may be justified. One reason is if the defender may not deserve abject privacy protections, such as when the defender is a criminal. Another reason is if a defender won't change a compromised password or private key, even after being told about the vulnerability. Telling that defender that the entire world will learn the password might be the drastic step needed to prompt the change. These examples, however, do not take away from the general point – it almost always helps the attackers more than the defenders to disclose the private key.

about vulnerabilities even after a high number of attacks. Where *L* is so low, the effectiveness of hidden surveillance persists.

To illustrate the difference between surveillance and most physical attacks, return to the example of the machine guns or the hidden pit covered by leaves. The attackers have a high *L* from these attacks – they learn about the location of the machine guns or of the existence of the hidden pit. By contrast, suppose that there are well-hidden observers or surveillance cameras that watch the attack. Even attacks that succeed quite possibly would not capture the observer or find out the strategy for the hidden cameras.

The same pattern exists for wiretaps or bugs on networked systems such as telephones or the Internet. A person using the telephone is not supposed to be able to tell if the line is tapped. Even a hacker who gets past a firewall may trigger alarms that the attacker can't perceive. Once again, there is a low *L* about surveillance defenses.

Those involved in surveillance have long understood the importance of preventing the opposition from knowing the nature of their surveillance. For instance, Neal Stephenson organizes his masterful novel *Cryptonomicon* around precisely this theme.[42] The novel retells the story of the Allies' decryption of the German Enigma encryption system during World War II. The strategic question for the Allies is how much to act on the secret messages they have decoded. For instance, if a convoy is crossing the Atlantic and the U-boats are poised to attack, should the convoy shift course? If the convoy does, then the Germans might deduce that the Enigma system has been broken, undermining the long-term ability to win the war. If it does not, then many people and boats will be lost. The novel describes elaborate efforts by the Allies to create cover stories for how they get useful intelligence. They seek to reduce the learning (*L*) by the attackers who are subject to surveillance.

The importance of retaining a hidden surveillance capability was also crucial to the entry of the United States into World War I.[43] At a time when Germany and the United States were officially neutral, the Germans sent the famous "Zimmerman telegram" to the government of Mexico. The telegram offered enticements for Mexico to ally with Germany against the United States, including promises of returning to Mexico territories that it held prior to the 1848 war. British intelligence decrypted the communication, but the intelligence agency was extremely loath to reveal to anyone else that it had the capability of breaking German codes. British intelligence then went through an elaborate, and

[42] NEAL STEPHENSON, CRYPTONOMICON (2002).

[43] The account here follows Singh*, supra* note 25.

successful, effort to make the leak appear to come from within the Mexican government. The Zimmerman telegram became public, speeding the entry of the United States into the war while retaining the British ability to conduct hidden surveillance of German communications.

These examples highlight the reasons that intelligence experts believe that sources and methods of surveillance should remain secret. They believe that disclosure of sources and methods will increase *L* and thus significantly help attackers. Even in many computer security settings, there is often a low *L* and surveillance measures can stay hidden. If these factual assertions are correct, as I believe they are, then disclosure of surveillance sources and methods will typically have a large help-the-attacker effect. Persons subject to wiretaps will stop talking on the phone. Persons who know that some radio frequencies are being monitored will shift to other frequencies, and so on.

It is vital to underscore the nature of the claim here: hiddenness about surveillance sources and methods will often improve security. This surveillance will improve the ability of defenders to protect their systems from the attackers. The claim is not, however, that hidden surveillance is therefore desirable (or lawful) in any particular setting. The assessment of overall desirability depends on judgments about multiple and often conflicting goals. Wiretaps and other surveillance, for instance, intrude on personal privacy. Fear of surveillance may chill desirable uses of communications networks, with negative effects on the economy and free speech. Public disclosure and debate about surveillance techniques are also crucial to holding government accountable. My current scholarly work on "The System of Foreign Intelligence Surveillance Law" examines these issues of security, privacy, and accountability in great detail.[44] The claim here is about the effectiveness of keeping surveillance hidden. Disclosure about sources and methods will rarely make the surveillance more effective at stopping attacks.

In summary, hidden surveillance techniques lead to a low level of learning (L) from attacks. Disclosure about the sources and methods of hidden surveillance is likely to reduce security, at least in the short term. Any overall judgment about the desirability of surveillance depends, in addition, on important other values such as the protection of privacy and the accountability of those conducting the surveillance.

[44] Peter P. Swire, *The System of Foreign Intelligence Surveillance Law*, __ GEO. WASH. L. REV. __ (forthcoming 2004).

5. When Do Attackers Already Know of the Vulnerability?

We now can return to the first assumption of the Open Source paradigm: that attackers will learn little or nothing from the disclosure. The discussion here has shown one setting in which attackers learn from the disclosure but in ways that benefit the defenders – attackers will sometimes be deterred when they learn about the defense. The discussion here has also identified two important categories where the assumption seems incorrect. First, private keys and the combinations to safes should stay secret, because the defense is based on the strategy of hiding that information from attackers. Second, surveillance techniques will often not be observable by the attackers, so the assumption that attackers already know about those techniques will often be wrong.

a. Discovering and Exploiting Vulnerabilities

With these important categories better understood, the main debate for software and network security concerns scenarios that do not primarily involve deterrence, private keys, or hidden surveillance. The main scenarios involve the following question: do hackers already know about, or will they promptly know about, the vulnerabilities in a firewall, a mass-market software program, or a cryptosystem?

In answering this question, an important variable is how hard it is for outsiders to discover a vulnerability. If it is generally easy to spot a vulnerability and exploit it, then the Open Source assumption will be correct – additional disclosure will not help the attackers much. Based on my discussions with computer security experts, however, there are at least three reasons to believe that spotting a new vulnerability in a mass-market software program is often more difficult. First, modern software programs often are incredibly complex, involving millions of lines of code.[45] Spotting any individual vulnerability requires considerable searching. There is thus often a lower number of attacks (*N*) on any piece of code than might otherwise be assumed. Second, the greater emphasis on computer security in recent years quite possibly has reduced the number of bugs per line of code. Third, my discussions with professional "bug hunters" suggest that finding a single vulnerability often takes considerable work by a highly skilled person. If one considers a cryptosystem rather than a mass-market software program, this last point is likely to be even more true – it will take a skilled person a considerable amount of work to find a vulnerability, if there is one.

[45] *See, e.g.*, Dennis Fisher, *Microsoft Puts Meat Behind Security Push*, EWEEK, (Sept. 30, 2002), *at* http://www.landfield.com/isn/mail-archive/2002/Oct/0004.html (discussing Microsoft's "massive bug hunt among millions of lines of its Windows code").

If a vulnerability is discovered, the next question is how difficult it is to write the exploit code to take advantage of the vulnerability. If it is hard to write the exploit code, then discovery of the vulnerability will not lead to rapid attacks on the vulnerability. This step may actually be easier for experts to do than one might have suspected. Based on my interviews with computer security experts, for large software programs an announcement about a flaw, even at a high level of generality, often quickly translates into exploit code. The "progress" from description of the vulnerability to successful attack happens due to high learning about what attacks work (*L*) and high communication with other potential attackers (*C*). Even a fairly general description of a vulnerability can focus skilled attackers on a subset of the millions of lines of code, speeding discovery of the exploit code.

The experts' rapid ability to exploit a vulnerability may or may not translate into non-experts' ability to do the same. At issue is the ability to attack by "script kiddies," the often-young hackers who find a step-by-step script on the Internet for doing an attack. Script kiddies can effectively attack a system that is configured in a standard way and has not installed a patch for a known vulnerability. On the other hand, defenders can often defeat script kiddies by altering the system in some way. In terms of the model developed in this article, uniqueness (*U*) by the defender aids the defense. Having a unique and hidden defense quite possibly will defeat attackers who are simply following a script.

b. The Analogy Between Exploiting Vulnerabilities and the Efficient Capital Markets Hypothesis

There is a useful analogy to the rich literature on the efficient capital market hypothesis (ECMH) in economics.[46] Efficiency in the ECMH means that the current price of the stock or other security accurately includes all the relevant information. Efficiency in the Open Source paradigm also means that all the relevant information is already known to outsiders – disclosure of a vulnerability does not help the attackers.

The claim here is that the Open Source paradigm has implicitly assumed what is called the "strong" form of the ECMH, that "current security prices fully reflect all currently existing information, whether publicly available or not."[47] The efficiency of capital markets, in this theory, depends on the actions of a large number of traders who follow

[46] Credit for the ECMH is often given to Eugene Fama, *The Behavior of Stock Market Prices*, 38 J. BUS. L. 34 (1965). For a detailed explanation of the ECMH in historical perspective, together with critiques of it, see Lawrence A. Cunningham, *From Random Walks to Chaotic Crashes: The Linear Genealogy of the Efficient Capital Market Hypothesis*, 62 GEO. WASH. L. REV. 546 (1994).

[47] *Id.* at 560.

the market extremely closely and exploit any opportunity to make a profit. The traders who analyze new information more quickly and accurately than other traders can gain a temporary advantage. The combined effect of all these traders is to push the market very quickly to the point where the price reflects all available information.[48]

The Open Source paradigm makes assumptions similar to the ECMH – attackers learn little or nothing from disclosure, because the attackers have already efficiently figured out the vulnerabilities. One can, however, identify some important differences. First, the number of traders in the capital markets is very high, with numerous experts and traders for even the lesser-known stocks. By contrast, there is not necessarily a supply of expert hackers for each aspect of the large computer programs and for each lesser-known computer programs. Second, the incentive structure to create "efficiency" is quite different. In the stock market, it is lawful trading that pushes the stock market to efficiency. The successful analyst buys stock and makes money immediately and in large quantities. By contrast, there is seldom a big cash reward for discovering a vulnerability (although the "bug finder" may develop a good reputation and gain consulting contracts). Exploiting the vulnerability also has different incentives – there are criminal penalties for attacking computers, so the incentive to use the knowledge is presumably lower than for lawful stock trading.

These differences would predict that the market for finding computer vulnerabilities is less efficient than the market for finding wrongly-priced securities. The likely inefficiency in finding vulnerabilities undermines the Open Source assumption that attackers already know about vulnerabilities or will promptly discover them. The likely inefficiency is even greater, moreover, in light of the criticisms made against the ECMH itself in recent years. There has been a significant and growing literature showing ways in which capital markets are not as efficient as the ECMH's proponents had previously thought.[49]

[48] The strong version of the ECMH assumes that the market price of the security reflects publicly-available information as well as information known only to the company itself. Under the semi-strong view, insiders might know additional information that would shift the price of the security if publicly revealed. This "insider information" can thus be valuable to insiders because they can predict the price better than public traders. Section 10(b) of the Securities Act of 1934 prohibits insider trading. 15 U.S.C. § 10(b) (2004).

The semi-strong view of the Open Source paradigm, by analogy, would state that insiders might know of vulnerabilities that are unknown to the outside attackers. The efficiency of the market would be determined by how well the outsiders could detect and exploit the vulnerabilities that do not depend on having such insider information.

[49] *See, e.g.*, William T. Allen, *Securities Markets as Social Products: The Pretty Efficient Capital Market Hypothesis*, 28 J. CORP. L. 551 (2003); Lynn A. Stout, *The Mechanisms of Market Inefficiency: An Introduction to the New Finance*, 28 J. CORP. L. 635 (2003).

Additional research might fruitfully develop the analogy further between the ECMH and the efficiency of finding vulnerabilities in computer systems. For instance, researchers might explore each of the criticisms of the ECMH in order to examine the possible sources of inefficiency in the exploitation of vulnerabilities. By analyzing the possible sources of inefficiency, computer security researchers can identify areas where vulnerabilities are less likely to be known to attackers, and where disclosure is thus more likely to provide substantial assistance to attackers.

On the other hand, there are scenarios where attackers have very strong incentives to discover vulnerabilities. In future military conflicts, for instance, attackers will be highly motivated to discover any vulnerability in the computer or network systems held by their enemy. Where there are determined and well-financed adversaries, the attackers may be very effective at discovering vulnerabilities. One can therefore imagine the following counter-intuitive situation. Civilian attackers may be inefficient, so that disclosure has a large help-the-attacker effect. Military attackers, by contrast, may be efficient in exploiting vulnerabilities that can be perceived from the outside. For those vulnerabilities, greater disclosure might actually be rational. The disclosure will do little or nothing to help the attackers, but there may be help-the-defender effects for system designers or for other defenders who rely on the system.

In summary, there is likely greater inefficiency today in the discovery of computer and network vulnerabilities than assumed in the Open Source paradigm. The analogy to the Efficient Capital Markets Hypothesis shows that the degree of efficiency depends on the incentives and institutional arrangements that attackers have to discover and communicate about vulnerabilities.

B. The Assumption that Disclosure Will Tend to Improve the Design of Defenses

The next assumption is the one most strongly held by Open Source proponents— disclosure of code and vulnerabilities will improve security because it will result in improved design of defenses. As firewall experts Chapman and Zwicky have written:

> Some people feel uncomfortable using software that's freely available on the Internet, particularly for security-critical applications. We feel that the advantages outweigh the disadvantages. You may not have the 'guarantees' offered by vendors, but you have the ability to inspect the source code and to share information with the large community that helps to maintain the software. In practice, vendors come and go, but the community endures.[50]

In this article, I do not take a position on the almost theological issue of whether Open Source software provides better security than proprietary software.[51] Instead, the discussion here seeks to identify some of the variables that would tilt the outcome in one direction or the other. The previous discussion showed how mass-market software and firewalls are subject to more efficient attacks due to the high number of attacks (*N*), learning from attacks (*L*), and communication among attackers (*C*). The focus here is on how defenders can alter the defenses (*A*) in ways that improve the defense over time. After looking at variables that affect when Open Source or proprietary software may provide better security, the discussion turns to how openness has particular value in promoting security and accountability in the long run.

1. Variables that Affect When Open Source or Proprietary Software May Provide Better Security

Consistent with the goals of this paper, the effort here is to identify situations where openness is more or less likely to improve security. In this discussion, it is helpful to distinguish between two meanings of "open." The focus of the discussion in this paper is on "open" in the sense of "not hidden." In particular, outsiders can generally see the source code for Open Source software but not for proprietary software. This paper does not address the extent to which software should be "open" in the sense of "not owned" under copyright or other laws.

[50] D. Brent Chapman & Elizabeth D. Zwicky, BUILDING INTERNET FIREWALLS 23 (1995).

[51] In the interests of full disclosure, I note that I am a member of Microsoft's Trustworthy Computing Academic Advisory Committee, which is a group of 19 academics that has been asked to provide advice on security and privacy issues to Microsoft. I have also discussed the issues in this paper at great length with many Open Source advocates. The views expressed herein are entirely my own.

a. Expertise of Inside and Outside Programmers

The security of proprietary software relies substantially on the expertise of "inside" programmers. These individuals are employees or contractors of the company that owns the mass-market software or the organization that operates the firewall.[52] By contrast, the Open Source paradigm relies on "outside" programmers – individuals who usually are not employed by or under contract with whomever initially designed the software.

The respective effectiveness of either the Open Source or the proprietary approaches will depend on the relative quantity and expertise of inside and outside programmers. Chapman and Zwicky emphasize the advantage of outside programmers when they refer to "the large community that maintains the software."[53] In other settings, an organization might be able to bring more and better programmers, who have the relevant expertise, to the inside. For example, consider the software for specialized military uses such as for launching rockets. In such a setting, there may not be a "large community [on the outside] that maintains the software."[54] The more effective approach, in the absence of that outside community, quite likely would be to rely on inside programmers—persons who are hired or trained by the military. In such instances, disclosure of the software does not enlist a community of outside programmers, although it may help the attackers find vulnerabilities.

b. The Incentives to Improve the Defense

One chief argument by supporters of the proprietary approach is that the owner of the software or the system has strong incentives to provide security. The reputation of the software manufacturer or system owner is on the line, and bad security can lead to a direct loss of revenue. In the Open Source model, the incentives are less clearly defined. The outside programmers might gain a good reputation by designing a patch. A good reputation might translate into consulting contracts or other remunerative work, although the time spent working on a patch seems less directly profitable for the Open Source programmer than it is for a company that increases sales due to better security. Open Source programmers may also improve software due to a combination of other

[52] Proprietary organizations may also get tips about problems and solutions from users of the software and other outsiders, but the emphasis is likely to be on inside programmers.

[53] Chapman & Zwicky, *supra* note 50, at 23.

[54] *Id.*

motives, including membership in a community of programmers and the feeling of satisfaction from helping solve the problems facing others.

The extent to which one approach—proprietary or Open Source—will provide greater incentives to improve the defense is essentially a question of sociology and organizational behavior. Over time the sociological context might shift. A vibrant Open Source community in one period might descend into a "what's in it for me" approach in a later period. Alternatively, a vibrant Open Source community might become broader and deeper in its skills over time compared with the inside programmers available to proprietary efforts.

c. Persistence of the Expertise

Chapman and Zwicky point out the risk that any single company can disappear: "In practice, vendors come and go, but the community endures." Users of a proprietary product thus risk the chance that the company will go bankrupt or otherwise stop supporting the product. On the other hand, there are scenarios where the proprietary approach would likely lead to better persistence of expertise. The owner of a software program or a firewall might invest in having specialized expertise.[55] For instance, the owner of a software program may find it worthwhile to keep on staff a person who is expert in one complex piece of a software program. Similarly, the military might decide to keep on staff persons who are experts in software that only the military uses. In these instances, the proprietary model may well create more persistent expertise than an Open Source approach.

d. The Institutional Context for Patching

The usual Open Source belief has been that patching – the release of improved code that addresses a vulnerability – works better where the Open Source community can probe for vulnerabilities and then fix them. The accompanying belief has been that many proprietary companies have been reluctant to admit to vulnerabilities or to invest the resources to issue good patches.

My interviews with computer security experts suggest that these conclusions have quite possibly become less true over time. First, proprietary companies have shifted to a norm of issuing patches as part of the overall effort to improve cybersecurity. Second, proprietary companies have in some instances created substantial institutional structures to create and disseminate patches. These institutional

[55] The economist Oliver Williamson calls this sort of investment "transaction specific capital" and considers it an important predictor of where firms make investments. OLIVER E. WILLIAMSON, THE ECONOMIC INSTITUTIONS OF CAPITALISM 30-32 (1998).

structures can respond to vulnerabilities in a coordinated way. A coordinated approach, if carried out effectively, may lead to faster and more consistent responses to problems across different platforms.

The point here is not to announce that one approach or the other is necessarily the clear winner when it comes to effective patching. Instead, the speed and quality of patching is likely to vary over time depending on the institutions that exist to create and disseminate patches.

e. Interoperability and Openness

Having Open Source code can facilitate interoperability. System owners will have the ability to see what they are including in their system and how to integrate it with existing programs. For this reason, greater disclosure can improve security to the extent it permits system owners to know their systems and avoid vulnerabilities.

On the proprietary side, there is the usual market incentive to provide effective solutions for clients. Software companies want their products to work well for a large range of users. They often design their products to inter-operate with other products, and they work with system integrators to create overall systems that work.[56] As with other factors discussed in this section, the extent to which the possible advantages of openness outweigh the possible advantages of vendors directly seeking to satisfy the market by increasing sales is an empirical question.

2. The Role of Disclosure in Creating Long-Run Security and Assuring Accountability

Much of the comparison thus far of the Open Source and proprietary approaches implicitly concerns short and medium-term security, such as which approach would typically create a better patch for a newly discovered vulnerability. An additional basis for disclosing information is to improve *long-run* security. Bruce Schneier, for instance, states that "public scrutiny is the only reliable way to improve security—be it of the nation's roads, bridges and ports or of our critically important computer networks."[57] The belief is that organizations that rely on secrets are unlikely, in the long-run, to update their security effectively. On this view, testing by outsiders is crucial to overcoming inertia within the organization. Even if secrecy masks vulnerabilities in

[56] For a discussion of the incentives for software and other manufacturers to promote interoperability, see Joseph Farrell & Philip Weiser, *Modularity, Vertical Integration, and Open Access Policies: Towards a Convergence of Antitrust and Regulation in the Internet Age*, 17 HARV. J.L. & TECH. 85, 97-104 (2003).

[57] Bruce Schneier, *Internet Shield: Secrecy and Security*, S.F. CHRON., Mar. 2, 2003 at D5.

the short-run, the secretive organization is laying the groundwork for a larger-scale problem in the long-run.

It is difficult to provide empirical tests for when secrecy leads to long-run failure to adapt and modernize. One can try to compensate for secrecy by creating institutions that inspect and challenge the status quo without disclosing secrets to the world. Organizations can hire "Tiger Teams"[58] and other sorts of outside experts to probe for weaknesses. Organizations can hire independent auditors and create oversight boards to watch for areas of weakness. The usefulness of these institutional responses will vary widely, but they are unlikely to be effective unless they can probe into possible areas of vulnerability and then have institutional support when they recommend changes.

Over the long run, the usefulness of openness for promoting security overlaps with the usefulness of openness for assuring accountability more generally. The Freedom of Information Act[59] and other openness mechanisms are useful in part because they allow vulnerabilities to be discovered and security to be improved. These mechanisms are also useful, however, for exposing corruption, abuse of power, and the other evils that can flourish in secret. It is a topic of my continuing research to shed light on situations where openness is most important to accountability and long-run improvement in security. For purposes of this article, however, the claim is more modest. Once one has done the analysis on the extent to which disclosure helps security, there is reason to place a thumb (and perhaps an entire palm) on the scale on the side of disclosure. That tilt is due to the recognition of the likely long-run decrease in security and accountability that comes from secrecy. The longer that information is designed to stay secret, the greater the risk to system security and general accountability.

On the comparison of Open Source and proprietary software, this article does not take a position on the contentious issue of which approach provides better overall security. Significant variables include: the relevant expertise of inside and outside programmers; the incentives to improve the defense; the persistence of relevant expertise; the institutional context for patching; and how interoperability is assured. Disclosure is often additionally useful for promoting long-run security and assuring accountability.

[58] Michael Lee et al., *Electronic Commerce, Hackers, and the Search for Legitimacy: A Regulatory Proposal*, 14 BERKELEY TECH. L.J. 839, 884 n. 195 (1999) (defining "Tiger Team" as computer security experts, hired by the owner of a computer system, who simulate hostile break-ins).

[59] 5 U.S.C. § 552 (2004).

C. The Assumption that Disclosure Will Spread Effective Defenses to Others

The third assumption in the Open Source paradigm is that disclosure will spread effective defenses to third parties. The assumption in the military paradigm is that disclosure will prompt little or no improvement in defense by other defenders. The discussion here will explain when each assumption is more likely to be true and thus when disclosure is likely to help other defenders.

The military assumption is more convincing in settings where mechanisms exist to disclose only to trusted defenders. In the military setting, there are strongly-authenticated fellow defenders, such as others that share the same uniform. When changes should be made in the defenses, the military has a hierarchical command structure. Orders can be given to the appropriate units to implement the change in defense. Under these assumptions of strong authentication and an established hierarchy, disclosure to the entire world has low or zero benefits (the other defenders improve their defenses) and potentially significant costs (the help-the-attacker effect).[60]

The situation changes for mass-market software. There is no strong authentication for who is an "authorized security specialist" for widely-used software. Suppose that a large software company tried to send security information to every "authorized security specialist" while trying to keep the information secret from all potential hackers. That sort of mass notification, with no leakage, is highly unlikely to succeed. In the absence of strong authentication that separates "good guys" from "bad guys," the disclosure that does occur will generally be available to both. The military option of selective disclosure is much less likely to be available.

The mass-market software programmer also has less hierarchical control over defenses than does a military commander. For mass-market software, many users lack expertise. Many defenders also may not pay much attention to the programmer's plea to install patches or otherwise upgrade the defense. Given the lack of hierarchical control, those seeking to spread the new defensive measure may have to rely on widespread publicity to alert the third-party defenders about the threat.

[60] In the real life of the military, of course, the assumptions of strong authentication and effective hierarchy do not always exist. Spies might learn about information that was supposed to transmit only to members of the military. Orders might not be followed. Nonetheless, the ability to get information to selected fellow defenders is likely much greater in a well-run military organization than it is for mass-market software companies.

In short, disclosure is more likely to help attackers where there is a unified defense (one organization with hierarchical controls). Disclosure is more likely to help defenders where there are numerous third parties that risk harm from a vulnerability. The latter situation is more likely to occur for the major settings for computer and network security. For firewalls and mass-market software there are many ordinary users who might have the vulnerability. For encryption, the messages of many users are subject to attack if the cryptosystem is broken. Because there are so many third parties, disclosure becomes more important in order to alert defenders about whether a product is secure or whether a patch is needed.

Interestingly, the needs of the U.S. military seem to have played a role in prompting mass-market software companies to disclose more about vulnerabilities. Over time, the military has followed the dictates of Congress and bought a great deal of commercial off-the-shelf (COTS) software. Based on my interviews with security experts, [61] the military became aware, over time, of software vulnerabilities that had not been disclosed either to the military or to the general public. The military found it unacceptable to be vulnerable to attacks that were known to attackers and to the software company, but not to the military. The military thus put pressure on software providers to increase the disclosure of vulnerabilities, so that users such as the military would be in a better position to know about vulnerabilities and develop a response.

In summary, disclosure will tend to help the attackers but not the defenders in a military setting, where there is strong authentication of defenders and an established hierarchy to implement better defenses. Disclosure provides greater benefit to defenders when there are numerous third-party users, no effective way to communicate only to friendly defenders, and no hierarchical way to ensure that defenses are put into place.

[61] Persons in both the public and private sector provided the information about this history to me as background.

CONCLUSION: SECURITY, PRIVACY, AND ACCOUNTABILITY

This paper has addressed when disclosure of information will improve security. "Security," for this paper's purposes, is defined as preventing the attacker from gaining control of a physical installation or computer system.[62]

There are clearly other compelling goals to consider in deciding when to disclose information. Accountability usually increases with greater disclosure. Disclosure of information can produce economic winners and losers. Free speech and other First Amendment issues are implicated by disclosure policy. Personal privacy, the subject of much of my previous academic and government work, can also be compromised when information is disclosed[63] Compelling goals such as accountability, economic growth, free speech, and privacy should be included in any overall decision about whether to disclose information.

An essential part of the analysis, however, is to understand when disclosure helps security itself. Understanding this is important in its own right, as an intellectual topic that not that has not received sufficient attention to date. It is also crucial in the debates about how to create cyber-security and physical security in the age of the Internet and of terrorist threats.

[62] A more expansive definition of "information security" is given in the Federal Information Security Management Act of 2002, Pub. L. 107-347, 116 Stat. 2946, Sec. 301:

> (1) The term 'information security' means protecting information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction in order to provide—
>
> > (A) integrity, which means guarding against improper information modification or destruction, and includes ensuring information nonrepudiation and authenticity;
> >
> > (B) confidentiality, which means preserving authorized restrictions on access and disclosure, including means for protecting personal privacy and proprietary information; and
> >
> > (C) availability, which means ensuring timely and reliable access to and use of information.

[63] From 1999 until early 2001 I served as the Clinton Administration's Chief Counselor for Privacy, in the U.S. Office of Management and Budget. This research project arises from my perception at that time that some of the hardest and least understood issues concerned the intersection of privacy and security. My work on the topic began in the summer of 1999, when the Federal Intrusion Detection Network ("FIDNet") became a topic of controversy. In the wake of criticism of FIDNet (see John Markoff, *U.S. Drawing Plan That Will Monitor Computer Systems*, N.Y. TIMES, July 28, 1999, at A1), I was asked to work with Richard Clarke's cyber-security team to ensure that federal computer security was done consistently with privacy and civil liberties. The next year, I served as chair of a White House Working Group addressing how to update electronic surveillance laws for the Internet Age, another topic where privacy and security concerns intersected. Since my return to academic life, much of my writing has addressed the intersection of security, privacy, and surveillance issues. My privacy and other publications are available at www.peterswire.net.

This paper seeks to correct two common misunderstandings about when disclosure improves security. Secrecy helps much more often than is suggested by the slogan "there is no security through obscurity." In presenting earlier versions of this article, the most sophisticated technologists have understood this fact. They have known that keys and passwords should remain secret and that a good firewall can benefit from idiosyncratic features that defeat the script kiddies. This paper draws on the literature about Efficient Capital Markets Hypothesis, however, to suggest that the efficiency of attackers in discovering vulnerabilities will often be less than Open Source proponents have presumed. More broadly, my discussions with security experts have uncovered no models or systematic ways to analyze the limits of what I have called the Open Source paradigm for security through openness.

The paper also teaches that disclosure improves security much more often than is suggested by the slogan "loose lips sink ships." First, military systems often rely on commercially-available software, and the security of those systems thus depends on military system owners learning about vulnerabilities. Second, military actions are subject to the growth of the Public Domain, where information gets communicated so quickly and effectively among potential attackers.[64] Third and most broadly, the model in this paper suggests that openness may be the best general strategy in situations with low uniqueness, where there are high values for number of attacks (N), learning from an individual attack (L), and communication among attackers (C).

In terms of methodology, this article has offered an economic analysis for determining when "there is no security through obscurity" and when "loose lips sink ships." The first step is to assess the costs and benefits of the disclosure with respect to potential attackers. In some instances, for strong positions, disclosure will deter attacks and is thus beneficial. In other instances, disclosure tends to spread information about vulnerabilities. Even then, where the facts fit the Open Source and Public Domain paradigms, disclosure will offer little or no aid to attackers. Thus, disclosure can go forward if there are benefits to the defenders or if other values favor disclosure. When the facts fit the Information Sharing and Military paradigms, disclosure is more likely to help the attackers. Nonetheless, disclosure is more likely than previously

[64] The growth of the Internet, with its lack of national boundaries on communications, has lowered the cost and increased the effectiveness of research about other countries. Military commanders expect to use new technologies to "see the battlespace" and have "integrated sight" of the enemy's capabilities. ADMIRAL BILL OWENS, LIFTING THE FOG OF WAR 119, 133 (2000) (describing greater information gathering and processing as a central part of the "Revolution in Military Affairs"). Opposing forces will similarly pursue strategies of high N, L, and C to "life the fog of war."

to have net benefits in cyber-settings and other modern settings where attackers can mount numerous attacks (high *N*), gather information cheaply (high *L*) and communicate efficiently about vulnerabilities (high *C*).

Another important theme of the article is that there are often third-party defenders who benefit from disclosure about vulnerabilities. The Military paradigm implicitly assumes that defenders act in a unified way with strong authentication (the ability to recognize allied soldiers) and hierarchical control (the ability to order fixes for vulnerabilities). When these assumptions no longer hold, such as for mass-market software and networks operated by diverse actors, then disclosure is much more likely to have net benefits for defenders.

By defining the factors that contribute to high uniqueness, this article identifies the variables that determine when secrecy will improve the defense: the effectiveness of the defensive feature against the initial attack (*E*); the number of attacks (*N*); the degree of learning by the attacker (*L*); the degree of communication with other potential attackers (*C*); and the extent to which defenders can effectively alter the feature before the next attack (both alteration by the system designer (*A-D*) and alteration by third parties, such as Open Source programmers (*A-T*)).

Identification of these variables provides the answer to the question asked in the paper's title: What is different about computer and network security? For key computer and network topics such as firewalls, mass-market software, and encryption, the effect of variables such as high *N*, *L*, *C*, and *A-T* show why the benefits of disclosure of vulnerabilities often outweigh the benefits of secrecy. Disclosure is not necessarily or logically more desirable for computer and network security than for physical security, but the crucial variables much more often result in having net benefits from disclosure. Examination of the variables also illuminates important special cases, such as why disclosure of passwords and private keys will almost always be harmful and why surveillance generally benefits from secrecy concerning sources and methods.

In closing, the intellectual structure of this paper provides a systematic way to identify the costs and benefits of disclosure for security. Further research can assess the empirical levels of the relevant variables in different security contexts. Additional study can enrich the theoretical structure for assessing the effects of disclosure on security, such as by drawing more on the Efficient Capital Markets Hypothesis literature to identify where vulnerabilities are most likely to be discovered by attackers. Finally, further research can better explain how security goals should be integrated with other compelling goals such as accountability, economic growth, free speech, and privacy.

VIDEO GAMES AND REVERSE ENGINEERING:

BEFORE AND AFTER THE DIGITAL MILLENNIUM COPYRIGHT ACT

JOE LINHOFF[*]

TABLE OF CONTENTS

[*] I would like to thank my classmate Judith Richards and professor Philip Weiser for their help and encouragement with this note.

INTRODUCTION

One way to learn is by taking things apart. Reverse engineering is the process of using tools to analyze a product, in a way its designer did not intend, to learn how it works.[1] Reverse engineering is a cornerstone of innovation and an intellectual property "safety valve."[2] Reverse engineering plays an essential role in keeping the video game industry healthy and competitive.

Even if you're not a video game fan, reverse engineering in the video game industry is important. The industry brings a large number of disciplines together into a single consumable package—the video game.[3] The industry has produced enormous revenue growth in a relatively short period of time.[4] This growth has occurred largely under the radar of big business and regulation as video games historically have been considered a novelty or entertainment aimed at a narrow subculture. However, advances in the industry have turned this novelty into mainstream entertainment with revenues comparable to movies.[5] Now the industry appears on a lot of radars including those of Sony,[6] Microsoft,[7] Qwest,[8]

[1] *See* Pamela Samuelson & Suzanne Scotchmer, *The Law and Economics of Reverse Engineering*, 111 YALE L.J. 1575, 1578 (2002) (defining reverse engineering as "the process of extracting know-how or knowledge from a human-made artifact"); Kewanee Oil Co. v. Bicron Corp., 416 U.S. 470, 476 (1974) (defining reverse engineering as "starting with the known product and working backward to divine the process which aided in its development or manufacture"). In *Sony Computer Entertainment v. Connectix Corporation,* the court defined reverse engineering as:

> Reverse engineering encompasses several methods of gaining access to the functional elements of a software program. They include: (1) reading about the program; (2) observing "the program in operation by using it on a computer;" (3) performing a "static examination of the individual computer instructions contained within the program;" and (4) performing a "dynamic examination of the individual computer instructions as the program is being run on a computer."

Sony Computer Entm't, Inc. v. Connectix Corp., 203 F.3d 596, 599 (9th Cir. 2000).

[2] Universal City Studios v. Reimerdes, 111 F. Supp. 2d 294, 322 (S.D.N.Y. 2000) (recognizing the doctrine of fair use as a "safety valve").

[3] Video games bring together software, hardware, music, sound effects, choreography, story telling, user interface design, physical simulation, database programming, communications, and more.

[4] *See* John Markoff, *Recession? Don't Tell The Video Game Industry*, N.Y. TIMES, May 24, 2002, at C4 ("Sales of game software alone reached $6.4 billion last year, putting the game industry in striking distance of Hollywood, which had box-office sales of $8.35 billion in 2001. And video game executives predict this year will be even stronger."); Chuck Salter, *Playing To Win*, FAST CO., Dec. 2002, at 80 ("Last year, U.S. computer- and video-game revenue surpassed domestic box-office receipts, and this year, the game industry is expected to widen that gap with more than $10 billion in sales.").

[5] Markoff*, supra* note 4.

[6] "Sony Computer Entertainment America Inc. (SCEA) markets the PlayStation family of products and develops, publishes, markets, and distributes software for the PS one

the Federal Communications Commission,[9] and Congress.[10] Reverse engineering is a technique widely used in the industry to understand and improve on others' work. It is also used to gain access to third party game machines, giving motivated game designers access to standard platforms. Reverse engineering plays an essential role in the industry's growth and is now threatened.

This note shows how reverse engineering is used in the video game industry and how the Digital Millennium Copyright Act (DMCA)[11] can make reverse engineering a crime. The note argues that Congress should amend the DMCA to expand allowances for reverse engineering practices. Section I provides the intellectual property (IP) context for reverse engineering. Section II explains how reverse engineering is used in the video game industry. Section III explores some of the pre-DMCA case law and the reverse engineering balance arising from those cases. Section IV looks at the DMCA and the impact the DMCA will likely have on reverse engineering in the video game industry. Section V concludes this examination by calling for Congress to amend the DMCA to allow reverse engineering practices.

console and the PlayStation2 computer entertainment system for the North American market." *See* PLAYSTATION, ABOUT SCEA, *at* http://www.us.playstation.com/about.aspx (last visited Jan. 6, 2004).

[7] Microsoft notes on its web site that one if its "seven core business units" is "Home and Entertainment, including Microsoft Xbox, consumer hardware and software, online games, and our TV platform." *See* MICROSOFT, OUR COMMITMENT TO OUR CUSTOMERS, THE BUSINESS OF MICROSOFT (Jan. 25, 2004), *at* http://www.microsoft.com/mscorp/articles/business.asp.

[8] Qwest is looking to video games to add to the demand for broadband Internet connections. *See* Qwest CEO Dick Notebaert, Remarks at the Silicon Flatirons Telecommunications Program at the University of Colorado School of Law: Cleaning Up the Telecom Mess (Feb. 26, 2003) [hereinafter *Notebaert Remarks*] (transcript available through the Silicon Flatirons Telecommunications Program at http://www.silicon-flatirons.org).

[9] "Broadband technology will potentially allow users to download more information, including new multimedia applications, streaming news, music, games . . . ." FCC, BROADBAND: FREQUENTLY ASKED QUESTIONS, *available at* http://www.fcc.gov/cgb/broadband.html (last visited Jan. 6, 2004).

[10] *Violence in the Media: Antitrust Implications of Self Regulation and Constitutionality of Government Action*, *hearing before the Senate Comm. on the Judiciary*, 106th Cong. (Sep. 20, 21 2000), *available at* http://frwebgate.access.gpo.gov/cgi-bin/useftp.cgi?IPaddress=162.140.64.88&filename=74413.wais&directory=/disk2/wais/data/106_senate_hearings.

[11] Digital Millennium Copyright Act, Pub. L. No. 105-304, 112 Stat 2860 (1998) [hereinafter DMCA].

I. THE INTELLECTUAL PROPERTY CONTEXT OF REVERSE ENGINEERING

The practice of reverse engineering is subject to different treatment by different IP regimes. For example, patent law does not provide a reverse engineering defense to infringement, whereas trade secret laws do allow reverse engineering. The status of reverse engineering in the copyright regime is questionable and the subject of this note. This part looks more closely at the IP context of reverse engineering.[12]

A. Intellectual Property Law and Reverse Engineering

Patent law does not directly address reverse engineering. However, patents can deter and endanger reverse engineering by making the results of the effort unusable.[13] Patent law grants exclusive rights to an inventor to make, use, and sell an invention for up to 20 years.[14] The grant is "nearly absolute, barring even those who independently develop the invention from practicing its art."[15] Patent law does not provide a reverse engineering defense—patent holders have the right to sue those who reverse engineer their invention.[16] Video games have many elements that could qualify for patent protection including elements of hardware, software, algorithms, and data structures.[17]

Another danger to the practice of reverse engineering is the possibility that patent holders may extend protection beyond their actual invention. For example, the patent holder's exclusive right to make or use an invention could prevent others from "using" that invention in any of the steps that are necessary for reverse engineering.[18] Thus, if a patented element is difficult to decouple from unpatented elements, the

[12] Also discussed in this part, contract law, through shrinkwrap or click-through licenses, can be a great deterrent to reverse engineering.

[13] Julie E. Cohen & Mark A. Lemley*, Patent Scope and Innovation in the Software Industry,* 89 CAL. L. REV. 1, 21 (2001).

[14] 35 U.S.C. § 154 (2000).

[15] ROBERT P. MERGES ET AL, INTELLECTUAL PROPERTY IN THE NEW TECHNOLOGICAL AGE 14 (3d ed. 2003).

[16] *Id.* at 197.

[17] For a short list of issued software-related patent types, see ROBERT P. MERGES ET AL, INTELLECTUAL PROPERTY IN THE NEW TECHNOLOGICAL AGE 1032 (2d ed. 2000). Since *Chakrabarty*, the patent office has taken seriously the statement that almost anything under the sun that is made by man is patentable, except for laws of nature, physical phenomenon, and abstract ideas. Diamond v. Chakrabarty, 447 U.S. 303, 310 (1980) (citations omitted).

[18] Patent law does not provide an intermediate copying allowance such as is found in copyright case law as discussed later in the note. Thus, a patent holder could characterize copying a program as running afoul of the holder's exclusive right to "make" and "use" their invention. Cohen & Lemley, *supra* note 13, at 26.

patent could be used to block access to, and thus block reverse engineering of the unpatented elements. In this way, patents could be used to stop reverse engineering and pose a threat to the video game industry.

Trade secret laws are primarily state law doctrines that protect "against the misappropriation of certain confidential information,"[19] provided reasonable steps have been taken to keep the information secret.[20] Trade secret laws do not prohibit "reverse engineering a legally obtained product to determine the secrets contained inside."[21] Nor do they affect those who independently discover or invent a product. Thus, trade secret law alone does not present an obstacle to video game reverse engineering.

Before the DMCA, copyright's fair use doctrine allowed for reverse engineering. Copyright protects "original works of authorship fixed in any tangible medium of expression"[22] but does not protect ideas, procedures, processes, or methods of operation.[23] Thus, although copyright protects the fixed source code and the fixed object code, it does not protect the functional aspects of a computer program. Copyright infringement actions can be brought against someone who makes literal copies of a program,[24] or who has access to a copyrighted program and makes a program that is substantially similar.[25] But copyright law does not prevent independent creation, nor does it protect functional elements. Basic copyright law supports the growth of the video game industry because it prohibits literal copying but does not prevent sharing of the underlying ideas.

The doctrine of fair use provides an important limit on copyright's strength. Regardless of how a copy is made, it is not infringement if the copy is a fair use "for purposes such as criticism, comment . . . or

[19] MERGES ET AL, *supra* note 15, at 22.

[20] *Id.*

[21] *Id.* at 23. However, trade secret law combined with restrictive licensing terms prohibiting reverse engineering can prevent reverse engineering. *See* Bowers v. Baystate Tech., Inc., 320 F.3d 1317, 1323 (Fed. Cir. 2003).

[22] 17 U.S.C. § 102(a) (2000).

[23] "In no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work." *Id.* at § 102(b).

[24] Atari Games Corp. v. Nintendo of Am., Inc., 975 F.2d 832, 837 (Fed. Cir. 1992) [hereinafter *Atari I*]. *Atari I* was a Federal Circuit decision affirming the Northern District of California's decision to preliminarily enjoin Atari from exploiting Nintendo's copyrighted computer programs. *See id.* at 835. After the Federal Circuit's ruling in *Atari I*, the case returned to the Norther District of California where Nintendo filed a Motion for Summary Judgement against Atari. *See* Atari Games Corp. v. Nintendo of Am., Inc., 30 U.S.P.Q.2d 1401 (N.D. Cal. 1993) [hereinafter *Atari II*].

[25] *Atari I*, 975 F.2d at 837.

research . . . ."[26] Fair use permits "public understanding and dissemination of the ideas, processes, and methods of operation in a work."[27] Fair use also "permits an individual in rightful possession of a copy of a work to undertake necessary efforts to understand the work's ideas, processes, and methods of operation."[28] In this way, fair use provides legal permission for reverse engineering copyrighted works.[29]

To counteract the fair use exception, video game vendors have begun to rely on shrinkwrap licenses in attempts to prohibit reverse engineering. For example, the popular and controversial game *Grand Theft Auto* includes a "Limited Software Warranty And License Agreement."[30] The license states that the "act of installing and/or otherwise using the software" constitutes agreement to be bound to the terms of the license.[31] The terms of the license expressly prohibit reverse engineering and any copying of the software not specifically allowed in the license.[32] However, the enforceability of these licenses, including the question of when state contract law can preempt federal copyright law, is unsettled.[33] If shrinkwrap licenses become enforceable and binding on use of software, there could be "no logical stopping point" as to what "limitations on copyright protection might be eliminated."[34] Because of the uncertainty and great potential of shrinkwrap licenses to change the legal landscape, the topic is beyond the scope of this note.

[26] 17 U.S.C. § 107 establishes the four factors courts must use in examining fair use. *See infra* text accompanying note 146.

[27] *Atari I,* 975 F.2d at 843.

[28] *Id.* at 842.

[29] *See* Philip J. Weiser, *The Internet, Innovation, and Intellectual Property Policy*, 103 COLUM. L. REV. 534, 551 (2003).

[30] ROCKSTAR GAMES, GRAND THEFT AUTO, VICE CITY, TOURIST GUIDE 24 ¶ 1 (2003).

[31] *Id.* at ¶ 1.

[32] *Id.* at ¶ 5. You are allowed to install the software on your computer and "keep the original disk(s) and/or CD-ROM [] only for backup or archival purposes." *Id.* at ¶ 4. The license claims that "The Software and Accompanying Materials are protected by the United States copyright law and applicable copyright laws and treaties throughout the world." *Id.* at ¶ 3.

[33] *See* Bowers v. Baystate Tech., Inc., 320 F.3d 1317, 1323 (Fed. Cir. 2003) (holding the shrinkwrap license prohibiting reverse engineering enforceable. "Under First Circuit law, the Copyright Act does not preempt or narrow the scope of Mr. Bowers' contract claim."); ProCD, Inc., v. Zeidenberg, 86 F.3d 1447, 1449 (7th Cir. 1996) (holding shrinkwrap licenses "enforceable unless their terms are objectionable on grounds applicable to contracts in general"); *but see* Vault Corp. v. Quaid Software Ltd., 847 F.2d 255, 269 (5th Cir. 1988) (holding license term unenforceable because the provision in Louisiana's law that allowed licenses prohibiting adaptation using "decompilation or disassembly" conflicts with and "'touches upon an area' of federal copyright law"). *Vault* recognized that a license restriction "against decompilation or disassembly is unenforceable." *Id.*

[34] *Bowers*, 320 F.3d at 1338 (Dyk, J., concurring in part, dissenting in part).

B. No Coherent Treatment of Reverse Engineering

Patent, copyright, and trade secret law each treat reverse engineering differently.[35] The same game software, or parts of it, may be simultaneously protected under one or more of these regimes. [36] The danger is that if reverse engineering is not protected under each regime, it will lose protection altogether.[37] Commentators "advocate a coherent treatment of reverse engineering across intellectual property law."[38]

In a recent article, Julie Cohen and Mark Lemley suggested the creation of a coherent reverse engineering policy.[39] Cohen and Lemley "advocate a limited right to reverse engineer patented computer programs to permit study of those programs and duplication of their unprotected elements."[40] Under their treatment, reverse engineering of software would be treated consistently under patent, trade secret, and copyright law.[41]

A more developed, coherent treatment of reverse engineering of platforms is presented by Philip Weiser. Weiser presents a "competitive platforms model"[42] that would allow or prohibit reverse engineering depending on market conditions and the purpose of the reverse engineering.[43] A game maker would be allowed to reverse engineer a platform for vertical access, i.e. "between a platform and a complimentary product."[44] But prohibited from reverse engineering a competitor's platform for horizontal access, i.e. "between rival platforms."[45] The model views reverse engineering as a corrective action that should be allowed only after two preconditions are met.[46] First, "to the extent it seems clear that a company lacks market power," that company should be permitted to use its IP rights to prevent reverse engineering for

[35] *See, e.g.,* Weiser, *supra* note 29, at 551.

[36] *Id.* at 553.

[37] *See* Cohen & Lemley, *supra* note 13, at 27.

[38] Weiser, *supra* note 29, at 553; *see* Cohen & Lemley, *supra* note 13, at 6.

[39] Cohen & Lemley, *supra* note 13, at 6.

[40] *Id.*

[41] *Id.* at 29.

[42] Weiser, *supra* note 29, at 537. Regulation of rival systems is the core concern of this model. *Id.* at 556. This model would not allow firms to clone inventions as that would undermine important investment incentives. *Id.*

[43] As market conditions change, so would the legal protection. "[W]hen a platform standard reaches or is headed for a dominant position in a market, intellectual property protection against reverse engineering should recede." *Id.* at 591.

[44] Weiser, *supra* note 29, at 591. This occurred in Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510 (9th Cir. 1992). *See also infra* Section III.

[45] Weiser, *supra* note 29, at 560. "[T]he Ninth Circuit should have accepted Sony's claim of infringement." *Id.* at 602. This occurred in Sony Computer Entm't, Inc. v. Connectix Corp., 203 F.3d 596, 599 (9th Cir. 2000). *See also infra* Section III.

[46] Weiser, *supra* note 29, at 594.

horizontal access.[47] Second, IP rights should be enforceable to block reverse engineering when the platform market is developing—the goal of which is to provide investment incentives[48] and produce Schumpeterian competition.[49] Then, after these preconditions are met, where it seems clear a single standard will emerge as dominant, reverse engineering should be allowed to trump IP rights.[50]

Both approaches recognize the pro-competitive gains reverse engineering provides and argue for allowance of the practice across the IP regimes.[51] Either approach presents an improvement over the current state of affairs.[52] The next section looks at reverse engineering in the video game industry.

II. REVERSE ENGINEERING IN THE VIDEO GAME INDUSTRY

The video game industry is an incredible success. It started in the early 1970s[53] and in 30 years has grown to run neck and neck with Hollywood's box office revenues.[54] Sixty percent of Americans play video games.[55] It is on the cutting edge of the high-tech industry, yet it did not crash with the rest.[56] The video game industry is being looked at to fill telecom's broadband pipes.[57] It has been subject to very little external regulation and has turned only occasionally to the legal system for help. No manufacturer or game developer has been able to monopolize the

[47] *Id.*

[48] *Id.* at 584.

[49] *Id.* at 593. A strong Schumpeterian view promotes the use of 'roadblocks' around which innovators must find a way if they are to get to the market. The belief is that "market power is temporary," and "monopolies are both acceptable and necessary to facilitate technological innovation." *Id.* at 576-77. Weiser rejects "pure Schumpeterian thinking" as a driver for IP policy. *Id.* at 581.

[50] *Id.* at 593.

[51] *Id.* at 600; Cohen & Lemley, *supra* note 13, at 22.

[52] Cohen and Lemley's approach may be difficult in practice as a reverse engineer disassembling the software for a game will find it hard to distinguish between protected and unprotected elements. Weiser's approach will be particularly hard to apply in the video game industry due to the industry's 5 year cyclical nature – reverse engineering would be allowed when a manufacturer has market power in years 3 and 4, but not in years 1, 2, or 5 when the system is not as popular. It is also not clear if either approach sufficiently clears the hurdles out of the way of an unsophisticated entrepreneur.

[53] STEVEN L. KENT, THE ULTIMATE HISTORY OF VIDEO GAMES 25 (2001). The Magnavox Odyssey, the first console game system, was released in 1972, and several competitors joined the market the next year.

[54] Tom Standage, *Games Get Serious*, THE ECONOMIST: THE WORLD IN 2003, Dec. 2002, at 104; Markoff, *supra* note 5.

[55] INTERACTIVE DIGITAL SOFTWARE ASS'N, QUICK FACTS ABOUT VIDEO GAME CONSOLES AND SOFTWARE, *at* http://www.idsa.com/consolefacts.html (last visited Mar. 5, 2004).

[56] Standage, *supra* note 54, at 104; Steve Alexander, *Video-Game Industry Hopes to Take Success Online*, MINNEAPOLIS STAR TRIB., Sept. 30, 2002, at 1D.

[57] *Notebaert Remarks*, *supra* note 8; Standage, *supra* note 54, at 104.

market.[58] Small teams come out of nowhere and create best-selling games.[59] As discussed below, reverse engineering plays a major role in the industry's success.

This section contains interviews with three industry veterans who discuss three kinds of reverse engineering. Part A discusses how Mike Schwartz successfully reverse engineered access to the Sega Genesis game platform for Electronic Arts at about the same time as the *Sega v. Accolade* case.[60] Second, in Part B, Mark Loffredo talks about the reverse engineering of video game hardware. And finally, in Part C, Will Carlin explains how a kind of reverse engineering can be used to analyze, understand, and build on other game designs.

A. Reverse Engineering Platform Access

Mike Schwartz worked at Electronic Arts (EA) and reverse engineered the Sega Genesis.[61] EA found Sega's licensing agreement onerous, and decided to reverse engineer the system.[62] EA setup a "clean room modeled after the PhoenixBIOS case."[63] A clean room is used to monitor and control information flow.[64] Schwartz had to be screened off from the rest of the company while he was exposed to Sega's copyrighted information. Schwartz worked in a room named "Chernobyl." This was the smoking room and the only room with a door.[65] Chernobyl was next to the kitchen, had a big ceiling fan to suck out the smoke, and a bunch

[58] Magnavox, Atari, Nintendo, Sega, and Sony have at different times led the console market.

[59] id Software, a four person company, created *Wolfenstein 3-D* in 1992, *DOOM* in 1993, and *Quake* in 1996. ID SOFTWARE, INC, ID SOFTWARE BACKGROUNDER, *at* http://www.idsoftware.com/business/history (last visited Feb. 6, 2004).

[60] Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510 (9th Cir. 1992).

[61] Telephone Interviews with Mike Schwartz (Feb. 18, 2003; Mar. 6, 2003) (notes on file with author). This reverse engineering effort was accomplished at the same time and for access to the same platform that was at the center of *Sega v. Accolade.*

[62] Schwartz, *supra* note 61. The licensing deal was that Sega would manufacture the game cartridges and sell them back to the licensee. On top of this, the licensee would pay to Sega a fixed amount per game sold.

[63] In the late 1970s the BIOS on the IBM was successfully reverse engineered and licensed. BIOS stands for Basic Input Output System. It is a set of routines that serve primarily as a low level software interface to the hardware. This reverse engineering allowed others, including Compaq, to make IBM compatible motherboards. This was a critical step for "open architecture" and the success of the PC, it "set the course of computing in general." *Id.*

[64] For example, Schwartz had his email monitored by attorneys. *Id.* A clean room process usually involves two figurative rooms. The room where the engineers are exposed to copyrighted material could be called the 'dirty' room. The engineers in this room work to produce a manual that does not contain any protected material. In the second figurative room, the 'clean' room, engineers use the manual freely to create games.

[65] *Id.*

of EA's trophies in it.[66] For the task he used his own home-built reverse engineering software and hardware tools.[67]

In Chernobyl, Schwartz wrote a manual describing the functional elements required to program games for the Genesis. In terms of what the manual could contain, "addresses of registers and what they did was OK, but no snippets of code."[68] This manual could then be used by anyone in the company since the manual did not contain any of Sega's copyrighted information. Lawyers reviewed all the information he prepared before anyone on the outside could look at it.[69] At one point, he "corrupted someone by accident. This person became 'dirty'"[70] and was disqualified from subsequent development because of their exposure to protected information. Working in the clean room, it took Schwartz a month to reverse engineer the Genesis.[71]

This process turned out great for EA. They went to Sega in Japan, showed that they had reverse engineered the system, and were able to negotiate a very favorable licensing agreement.[72]

B. Reverse Engineering Hardware

Mark Loffredo has been designing hardware since 1982 and designed the arcade game hardware, including custom graphics chips, for many top selling games including *Mortal Kombat*, *Terminator 2*, *NBA Jam*, and *Crusin' USA*.[73] Reverse engineering arcade game hardware is not unheard of.[74] In the 1980s, there was a lot of paranoia in the industry that pirates were going to reverse engineer boards and make

[66] *Id.*

[67] *Id.*

[68] Schwartz, *supra* note 61.

[69] *Id.*

[70] *Id.* The purpose of the clean room was to keep Sega's copyrighted information from the rest of the employees since copyright does not protect independent invention, if there were ever to arise a question, EA could show that none of their employees, except Schwartz, ever had access to Sega's copyrighted work.

[71] It turned out the Genesis's graphics chip was very similar to one in the Colecovision system that he was familiar with. This saved him an enormous amount of time. He went home, got the manual for the other chip, and was able to test for differences. *Id.*

[72] *Id.* As far as Schwartz can remember, "Sega demanded to make all the carts. EA could buy as many carts from Sega as it wished, but had to pay Sega's price. Sega's price included something like $15/cart in usury fee!" *Id.* Schwartz, on the other hand, was not allowed to develop original Genesis games for fear that he would unconsciously repeat five lines of code. *Id.*

[73] Telephone Interview with Mark Loffredo (Mar. 6, 2003) (notes on file with author). *Mortal Kombat*, *Terminator 2*, *NBA Jam*, and *Cruisn' USA* are all arcade games developed and manufactured at Midway Manufacturing Co.

[74] In the 1981 case of *Midway Manufacturing v. Dirkschneider* the defendants were "engaged in the manufacture, distribution, and sale of video games…virtually identical to" *Galaxian*, *Pac-Man*, and *Rally-X*. 543 F. Supp. 466, 472 (D. Neb. 1981).

copies.[75] However, the instances of this actually happening were very low.[76] Complexity of new chips, encryption, and copy-protection advances make reverse engineering cost prohibitive.[77] "You really need to make a minimum number of games to make it worth while. This minimum would have to be in the multiple thousands."[78]

Loffredo "rarely looks at other chips" as he does not want "their design bias."[79] He would rather figure out something on his own than "shoe-horn an inadequate design that's hard to understand" into his system.[80] He is not sure exactly what is and what is not reverse engineering, but "ripping off concepts in the industry is widespread."[81] Everyone "continually looks at the competition and figures out how to do it better."[82]

Loffredo's latest system design is built around Xilinx[83] programmable logic chips.[84] These chips blur the line between hardware and software. They are programmed with a bitstream of data that is sent to the chip.[85] The data, which is a type of object code, defines how the chip acts. For example, a bitstream of code could program the chip to act like the 6502 microprocessor found in the early Apple II computers, then a different bitstream could be sent to the same chip to re-program it to act like the 68000 microprocessor found in early Apple Macintosh computers, and then a third bitstream of code could again be sent to re-program the chip to emulate the game of *Pong*.[86] In effect, software programs the hardware to be hardware. "It takes a lot of time to create the IP to program these chips."[87] Hardware engineers now create

[75] Loffredo, *supra* note 73.

[76] *Id.* There are supposedly over 90,000 *Robotron* games worldwide and only 60,000 manufactured legally – however, there was no copy-protection built into these games and they were relatively easy to copy. *Id.*

[77] *Id.*

[78] *Id.*

[79] *Id.*

[80] *Id.*

[81] Loffredo, *supra* note 73.

[82] *Id.*

[83] Xilinx, Inc., http://www.xilinx.com (last visited Jun. 22, 2004).

[84] Loffredo, *supra* note 73. Programmable logic devices, in contrast to fixed logic devices, allow the device's function to be programmed or reprogrammed at any time. XILINX, WHAT IS PROGRAMMABLE LOGIC?, *at* http://www.xilinx.com/company/about/programmable.html (last visited Jun. 21, 2004).

[85] Loffredo, *supra* note 73.

[86] *See* OPENCORES.ORG, *T65 CPU: Overview*, *at* http://www.opencores.org/projects.cgi/web/t65/overview (last visited Feb. 24, 2004). The Verilog for a Pong-like game is available at http://www.fpga4fun.com/PongGame.html (last visited Feb. 25, 2004).

[87] Loffredo, *supra* note 73.

hardware using programming languages and tools much like software engineer's tools and languages.[88]

This illuminates a fundamental issue as to the reach and importance of copyright law—the distinction between hardware and software is collapsing, and as copyright reaches to protect software, it also protects hardware.

C. Reverse Engineering Game Designs

Game designs are often based on, and evolve from, other games. *Asteroids* followed after and improved on *Space War, Galaga* improved on *Space Invaders, Mortal Kombat* improved on *Street Fighter*, etc.[89] Reverse engineering can be used by game designers to analyze and understand how a game is put together and what makes it work. Reverse engineering can uncover the internal rules of a game, how the scoring works, the pacing of the game, how the camera works, the game physics, the number of frames and timing of animations, and more. An experienced designer can determine some of these things by playing the game. However, a form of reverse engineering is useful for discovering other elements. The game designer's main reverse engineering tool is the video recorder. A game's inner-workings can often be discovered by watching a video of the game frame by frame, slowing the action down enough to see every detail and every change.

Game designer Will Carlin says "the whole industry is built on reverse engineering."[90] He started designing games in 1984. His latest game, *Big Buck Hunter*, has been the number one arcade game for two years.[91] Successful games are made by borrowing ideas.[92] For example, Sega's new game *Getaway* has the same play mechanics as Rockstar Games's *Grand Theft Auto*.[93] A game's play mechanics are the subtle combination of algorithms, math, physics, and ad-hoc programming that

[88] Hardware engineers can now "design a very complex logic circuit in front of a text editor." *Id.* They can design hardware using Verilog, a type of High-level Design Language (HDL) code, which "takes much of its syntax from the C language." *Id.* The HDL compilers and tools act very much like those for "software programming." *Id.*

[89] In *Space War, 1961*, the first real computer game, two ships flew around space much like those in Atari's 1979 game Asteroids. RUSEL DEMARIA & JOHNNY L. WILSON, HIGH SCORE 12, 49 (2002). Namco's 1981 game of *Galaga* improved on the basic design of Taito's 1978 game *Space Invaders*. *Id.* at 46, 76. Midway's 1992 game *Mortal Kombat* built on the design of Capcom's 1987 game *Street Fighter* which can trace its roots back to Data East's 1984 game *Karate Champ*. *Id.* at 280-81. *See Killer List of Video Games*, INTERNATIONAL ARCADE MUSEUM, *available at* http://www.arcade-museum.com (last visited Feb. 3, 2004) (provides a reference of arcade games and manufacturers).

[90] Telephone Interview with Will Carlin (Feb. 27, 2003) (notes on file with author).

[91] *Id.*

[92] Sometimes "it's downright plagiarism." *Id.*

[93] *Id.*

determine how the game responds to the controls. Carlin says a designer could video tape a car skidding around a corner and then analyze the skid marks in the video to get an idea of the game physics involved and how to recreate that effect.[94]

Video game software engineers, hardware engineers, and game designers all engage in different types of reverse engineering. Reverse engineering is important at each level. The industry has grown in leaps and bounds in large part due to competitor's' ability to understand, analyze, and build on other's work. If reverse engineering is outlawed, we should expect a big drop-off in the number of new games produced every year.[95]

III. PRE-DMCA CASE LAW FOR REVERSE ENGINEERING

The case law involving reverse engineering in the video game industry has focused primarily on copyright issues. This section examines three of the primary pre-DMCA cases that address reverse engineering of video game platforms.

A. *Atari v. Nintendo*

The controversy between Atari and Nintendo lays out most of the pre-DMCA framework for the legal analysis of reverse engineering in the video game industry. In the late 1980s, the 8-bit Nintendo Entertainment System (NES) had an 80% market share.[96] The security mechanism on the NES, called 10NES, prevented games from running on the system unless they contained a special chip and software.[97] Nintendo used the security mechanism to push game developers into licensing contracts.[98] Atari began reverse engineering the NES in 1986, the same year it was introduced in the US. [99]

Atari first tried to reverse engineer the security mechanism by "monitoring the communication" between the game cartridge and the game console.[100] However, this approach did not give them enough information. Next, in an attempt to re-create a listing of the object code, Atari "chemically peeled layers from the NES chips to allow microscopic examination of the object code."[101] However, this too failed as Atari's

[94] *Id.*

[95] Carlin, *supra* note 90.

[96] Atari Games Corp. v. Nintendo of Am., Inc., 897 F.2d 1572, 1574 (Fed. Cir. 1990).

[97] KENT, *supra* note 53, at 372.

[98] *Atari I*, 975 F.2d at 836-37 (Fed. Cir. 1992).

[99] KENT, *supra* note 53, at xiv. *Atari I*, 975 F.2d at 836.

[100] *Atari I*, 975 F.2d at 836.

[101] *Id.* Microscopic examination will not literally reveal the object code. *See* KENT, *supra* note 53, at 372.

engineers were not able to sufficiently reconstruct the code from the peeled layers of the chips.[102] Atari finally turned to their lawyers. As part of the copyright process, Nintendo had filed a listing of their object code with the Copyright Office. This listing contained the information Atari had unsuccessfully sought through reverse engineering. Atari's lawyers made up a fictional lawsuit, claimed that Nintendo was suing them for copyright infringement, submitted false affidavits to the Copyright Office, and got a copy of the listing.[103]

Atari was soon thereafter successful.[104] Atari developed its own security chip and program, which they named Rabbit, to mimic the 10NES.[105] In 1988, they began producing their own games "without Nintendo's strict license conditions."[106]

Nintendo and Atari sued each other. One of the issues was Atari's right to reverse engineer Nintendo's security mechanism. The court stated that except for the taint from their purloined copy of the 10NES program, Atari's reverse engineering was a fair use in so far as it was necessary to understand the 10NES.[107] "When the nature of a work requires intermediate copying to understand the ideas and processes in a copyrighted work, that nature supports a fair use for intermediate copying. Thus, reverse engineering object code to discern the unprotectable ideas in a computer program is a fair use."[108] However, because their copy of the 10NES program was fraudulently obtained, Atari lost this defense.[109]

Another related issue was whether Atari could copy program code that was not currently needed, but that might be needed in the future if Nintendo upgraded their security.[110] The court refused to extend fair use to a preemptive right to copy.[111] Atari further argued that the signal stream itself was not copyrightable.[112] Here the district court agreed, and found that the signal stream did not overcome the originality

[102] *Atari I,* 975 F.2d at 836.

[103] *Id.*; KENT, *supra* note 53, at 373.

[104] It is not clear how useful the stolen information was. *See* KENT, *supra* note 53, at 373 (discussing how Atari's clean room operation was close to breaking the 10NES at this time, and also Ed Logg's quote implying that no one used the information from the Copyright Office and merely that "some paralegal f---ed up!").

[105] *Atari I*, 975 F.2d at 836 (Fed. Cir. 1992).

[106] *Id.* at 836-37.

[107] *Id.* at 843.

[108] *Id.*

[109] *Id.* ("To invoke the fair use exception, an individual must possess an authorized copy of a literary work.").

[110] *Atari II*, 30 U.S.P.Q.2d at 1406-07.

[111] *Id.* at 1407.

[112] *Id.* at 1403.

requirement.[113] The ruling that the signal stream was not itself copyrightable would limit "Nintendo's rights in the 10NES program in [] two ways."[114] First, Atari, as a competitor, may copy those portions of the program that are necessary to access the unprotected signal stream for interoperability, and may include that code in their final version.[115] Second, they may make intermediate copies of the entire program in order to reverse engineer necessary sequences of the unprotected signal stream.[116]

However, the favorable ruling regarding the signal stream was not enough to overcome Atari's fraud.[117] Atari lost the dispute.[118] As *Atari v. Nintendo* was winding down, *Sega v. Accolade*, which addressed some of the same issues, was just beginning.

B. Sega v. Accolade

Accolade used a two-step clean room process to create video games compatible with the Sega Genesis game console.[119] The first step was to reverse engineer the system and create a development manual. Accolade purchased a Genesis video game console and three game cartridges.[120] Then they wired up the system so they could examine the data moving between the cartridge and the console during game play.[121] The engineers dumped the code from the cartridges, disassembled it, printed it, and studied it.[122] The engineers then loaded a mix of their own code and modified code from the purchased cartridges onto the console and tested it until they discovered how to unlock the Genesis.[123] "At the end of the reverse engineering process, Accolade created a development manual that incorporated the information it had discovered about the requirements for a Genesis-compatible game."[124] The manual did not contain any Sega code, but only contained "functional descriptions of the

[113] *Id.* at 1405 (citing Feist Publ'ns v. Rural Tel. Serv. Co., 111 S. Ct. 1282, 1289 (1991)).

[114] *Atari II*, 30 U.S.P.Q.2d. at 1408-09.

[115] *Id.* at 1408-09.

[116] *Id.*

[117] "To the extent, however, Nintendo is likely to show misappropriation and copying of the unauthorized Copyright Office copy, it is likely to succeed on the merits of its infringement claim." *Atari I*, 975 F.2d at 836 (Fed. Cir. 1992).

[118] The court granted Nintendo's summary judgment motion regarding the copyright infringement claim, finding elements in Atari's Rabbit program "firmly establish illicit copying." *Atari II*, 30 U.S.P.Q.2d at 1406-07.

[119] Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d at 1514 (9th Cir. 1992).

[120] *Id.* at 1514-15.

[121] *Id.*

[122] *Id.* at 1515.

[123] *Id.*

[124] *Id.*

interface requirements."[125]

The second step was to use the development manual to create its own games for the Genesis.[126] In 1990, Accolade released *Ishido*, a game that it had developed and released for the Macintosh and IBM PC.[127] In 1991, Sega began manufacturing a new version of the Genesis console with which *Ishido* would not work.[128] Accolade embarked on a second round of reverse engineering. The engineers found a small piece of code that was ignored by the original Genesis, but which was necessary to unlock the new Genesis.[129]

Sega filed a claim of copyright infringement against Accolade, not for the resulting product, but for Accolade's intermediate copying during their reverse engineering process.[130] The district court found for Sega primarily because Accolade's use was commercial—Sega had lost sales—and Accolade apparently had an alternative that did not require the intermediate copying of code.[131] The district court ordered Accolade to recall all of its infringing games within 10 business days.[132]

On appeal Accolade made four arguments relating to the copyright infringement claim: (1) intermediate copying is not infringement; (2) disassembly of object code to gain understanding of the ideas and functional concepts is lawful; (3) disassembly is authorized by section 117 which allows computer programs to be read into memory; and (4) disassembly in order to gain understanding of ideas and functional concepts is a fair use.[133] The court dismissed the first three arguments,[134] but accepted the fourth and dissolved the district court's order:[135]

> [D]isassembly of copyrighted object code is, as a matter of law, a fair use of the copyrighted work if such disassembly provides the only means of access to those elements of the code that are not protected by copyright and the copier has a legitimate reason for seeking such access.[136]

[125] Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d at 1515 (9th Cir. 1992).

[126] *Id.*

[127] *Id.*

[128] *Id.* The new console had an updated security system.

[129] *Id.* at 1515-16.

[130] *Id.* at 1516.

[131] Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d at 1517 (9th Cir. 1992). Sega claimed that there was an alternative way to make interoperable cartridges and was willing to show Accolade's attorneys, but not Accolade's engineers, the cartridges that accomplished this.

[132] *Id.*

[133] *Id.* at 1517-18.

[134] *Id.* at 1519 (noting that "intermediate copying of computer object code may infringe . . . regardless of whether the end product of the copying also infringes . . .").

[135] Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d at 1518 (9th Cir. 1992).

[136] *Id.*

This decision, along with *Atari v. Nintendo,* validated reverse engineering as a fair use defense. Fair use is explored further in the following case.

C. *Sony v. Connectix*

Sony v. Connectix took a closer look at fair use and addressed how reverse engineering could be used by a competitor[137] to create a compatible platform. Sony made the PlayStation video game console as well as games for the console.[138] PlayStation games were released on standard compact disks.[139] Connectix made a program, the Virtual Game Station (VGS), for the Apple Macintosh computer that allowed PlayStation games to be played on the Macintosh "even if you don't yet have a Sony PlayStation console."[140]

In the process of creating the VGS, Connectix analyzed the BIOS of the PlayStation.[141] The BIOS is a copyrighted program that acts as a low level interface between the software and the hardware.[142] Connectix engineers used the copyrighted BIOS for reference and testing only—none of the copyrighted BIOS appeared in the final VGS product.[143] The court found that Connectix's use of the copyrighted BIOS was a fair use.[144]

The "fair use doctrine preserves public access to the ideas and functional elements embedded in copyrighted computer software programs."[145] In determining whether a use qualifies under the doctrine, the Copyright Act lists the factors for consideration:

> (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and (4) the effect of the use upon the potential market for or value of the copyrighted work.[146]

[137] Sony Computer Entm't, Inc. v. Connectix Corp., 203 F.3d 596, 599 (9th Cir. 2000). Connectix and Sony were not ordinary competitors in the platform market. Connectix provided an alternate PlayStation-compatible platform that may have extended the sales of PlayStation games and thus increased the value of the PlayStation to Sony.

[138] *Id.*

[139] *Id.*

[140] *Id.* at 599, 601. They were also working on a Microsoft Windows version.

[141] Sony Computer Entm't, Inc. v. Connectix Corp., 203 F.3d at 601 (9th Cir. 2000).

[142] *Id.* at 599-600.

[143] *Id.* at 600.

[144] *Id.* at 602.

[145] *Id.* at 603.

[146] Sony Computer Entm't, Inc. v. Connectix Corp., 203 F.3d at 602 (9th Cir. 2000) (quoting 17 U.S.C. § 107 (2000)).

The first factor the court looked to was the nature of the copyrighted work. Because Sony's BIOS contained unprotectable elements that could not be examined without copying, the BIOS was accorded a "lower degree of protection than more traditional literary works."[147] Since the copying was necessary to examine those unprotectable elements, factor (2), the nature of the copyrighted work, weighed heavily in favor of Connectix.[148] As to the other factors, because the final product did not itself contain infringing material, factor (3), amount and substantiality, contributed very little to the analysis.[149] Factor (1), purpose and character, weighed in favor of Connectix because their product was "modestly transformative," did not merely supplant the PlayStation, and was a "wholly new product" notwithstanding the similarity of uses and functions.[150] Finally, because the VGS was "a legitimate competitor in the market for platforms on which Sony and Sony-licensed games [could] be played," factor (4), any economic loss incurred by Sony, "[did] not compel a finding of no fair use."[151]

These three cases show how reverse engineering acts as a fair-use balance to copyright law. The *Connectix* decision was published February 10, 2000. The new anti-circumvention rules of the DMCA went into effect a few months later on October 28, 2000.[152] Before Section IV examines how the DMCA alters the balance, the following part looks at the pre-DMCA reverse engineering balance.

D. The Reverse Engineering Balance

The above pre-DMCA decisions established a balance between a copyright holder's right to exclude, and a third party's right to access work through reverse engineering. This part looks closer at the balance focusing on two important factors: first, that manufacturers control the cost of reverse engineering; and second, that reverse engineering promotes competition by opening up access to platforms.

[147] *Id.* at 603 (quoting Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d at 1514 (9th Cir. 1992)).

[148] *Id.*

[149] *Id.* at 606.

[150] *Id.* at 606-07.

[151] *Id.* at 607.

[152] 17 U.S.C. § 1201(a)(1)(A) (2000) ("The prohibition contained in the preceding sentence shall take effect at the end of the 2-year period beginning on the date of the enactment of this chapter [enacted Oct. 28, 1998].").

1. Manufacturers Control Reverse Engineering Cost

Manufacturers control how difficult their platform is to reverse engineer. In designing the 8-bit NES, Nintendo used a combination of a hardware chip in each game cartridge and software embedded in their platform as a key and lock. This combination of hardware and software, largely because it included a hardware element, presented a significant access barrier to Atari. Compare this to the key and lock Sega used in their system: Sega used a generic software key, included in every game cartridge, to unlock the Genesis platform.[153] Sega's lock and key did not require an additional hardware chip in each game cartridge, as did Nintendo's system, and most likely was less costly to both Sega to manufacture and a competitor to reverse engineer.

Manufacturers control the cost of reverse engineering, both in terms of their own design and manufacturing costs, and also in terms of how difficult and costly the system is to reverse engineer. As the complexity of the lock and key grows, so does the effort required to successfully reverse engineer a system. Because manufacturers have the ability to control access to their platform without help from the legal system, legal protection can be redundant.

2. Platform Access

Platform access is often at the focal point of the reverse engineering debate.[154] Reverse engineering gives the third party a choice: negotiate with the platform manufacturer, or attempt to reverse engineer access. In addition to promoting fair licensing by placing an upper limit on the terms of acceptable licenses, i.e. the cost of reverse engineering,[155] this choice opens up access and markets for small developers without the resources or sophistication needed to negotiate a deal. This type of vertical access plays an important role in the industry's growth. Without access to standard platforms, small entrepreneurs are not able to participate in a large segment of the video game industry. For these

[153] Sega's system also triggers Sega's trademark. Their lock and key was named the Trademark Security System. Sega appears to have weighed the tradeoffs, and made a decision to implement a security system with low up-front costs that depended ultimately on trademark law.

[154] Vertical platform access was at issue in *Atari II* and *Sega v. Accolade*. Horizontal platform access was at issue in *Sony v. Connectix*. In the video game industry, horizontal platform access merits less discussion because of two important factors. First, as Judge Fern Smith notes in *Atari II*, there is a significant time lag needed to successfully reverse engineer a system. *Atari II*, 30 U.S.P.Q.2d 1401. And second, emulation of a system will almost always require next generation technology. Both of these factors give the manufacturer of a successful platform time to recover investments.

[155] *See* Weiser, *supra* note 29, at 548.

entrepreneurs, to the extent the law blocks reverse engineering, the law blocks access to the market. Losing this group of entrepreneurs will likely lead to an industry-wide loss of creativity, game content diversity, innovation, and competition. Because manufacturers themselves have the ability to dial in their own level of protection, redundant legal protections should be added carefully, lest they deter competition.

3. Healthy Balance

The reverse engineering balance before the DMCA was healthy. Pamela Samuelson and Suzanne Scotchmer studied the software industry generally, and concluded that reverse engineering had not hurt the industry.[156] Because "decompilation and disassembly are time-consuming and resource-intensive, these forms of reverse engineering [have] not . . . significantly undermine[d] incentives to invest in platforms."[157]

The courts also recognized that a balance was needed. Without the reverse engineering allowance, a copyright holder's rights are similar to those rights granted by the more stringent patent process. If a competitor "wishes to obtain a lawful monopoly on the functional concepts in its software, it must satisfy the more stringent standards of the patent laws."[158] Judge Fern M. Smith states the doctrine for allowing reverse engineering:

> By requiring independent game developers to carefully study a particular security system and discern which program instructions are truly necessary for present compatibility, console manufacturers will have a limited period of time in which to control the market for compatible games. In this time period, some third party game developers are likely to enter license agreements with Nintendo, particularly if they have limited resources. After a relatively short period of time, however, other developers will enter the game market with independently produced, but still compatible games. In addition, if third party developers who entered license agreements later find the license agreements too onerous, there still exists the option of reverse engineering the security system after the expiration of their license agreement. Thus, a fair use defense which allows copying for present compatibility balances the incentives for both game developers and console manufacturers.[159]

[156] Samuelson & Scotchmer, *supra* note 1, at 1612-13.

[157] *Id.* at 1622.

[158] Sony Computer Entm't, Inc. v. Connectix Corp., 203 F.3d at 605 (9th Cir. 2000) (citing Bonito Boats, Inc. v. Thunder Craft Boats, Inc., 489 U.S. 141 (1989)).

[159] *Atari II*, 30 U.S.P.Q.2d at 1407.

The three cases described above, *Atari v. Nintendo*, *Sega v. Accolade*, and *Sony v. Connectix*, address reverse engineering platform access and arrive at a healthy, competitive balance that allows some access while protecting incentives for manufacturers. The next section discusses how the DMCA throws off this balance.

IV. THE DMCA AND REVERSE ENGINEERING

The DMCA was signed into law October 28, 1998.[160] In addition to implementing the World Intellectual Property Organization Copyright Treaty,[161] the DMCA was enacted to support the "adaptation of the law of copyright to the digital age."[162] The DMCA significantly changes the law relating to reverse engineering and throws off the balance between a copyright holder's rights, fair use, and competition. The anti-circumvention provisions of the DMCA provide a way to seal off technology by severely restricting access through reverse engineering. This section looks at the text of the DMCA in light of its early judicial interpretations and discusses how it will likely affect the video game industry.

A. Anti-Circumvention and Reverse Engineering

In the digital age, reverse engineering and circumvention are two sides of the same coin. To the extent the DMCA prevents and limits circumvention, it prevents and limits reverse engineering. The anti-circumvention provisions of the DMCA prohibit the circumvention of a "technological measure that effectively controls access to a work protected under this title."[163] The effect is that any digital work protected under copyright has a new form of legal protection: anti-circumvention.[164] Section 1201 limits anti-circumvention in two important ways.[165] First, as stated in section 1201(a)(1)(A), anti-circumvention actions are limited to works that are "protected under this title."[166] This limitation, in conjunction with section 102, prevents anti-

[160] Digital Millennium Copyright Act, Pub. L. No. 105-304, 112 Stat 2860 (1998).

[161] *Id.*

[162] Universal City Studios v. Reimerdes, 111 F. Supp. 2d 294, 316 (S.D.N.Y. 2000).

[163] 17 U.S.C. § 1201(a)(1)(A) (2000).

[164] As long as a "technological measure" can be added to "control access" – this is a trivially low bar.

[165] 17 U.S.C. § 1201(c)(1) states that "[n]othing in this section shall affect rights, remedies, limitations, or defenses to copyright infringement, including fair use, under this title." If, as has been done in *Reimerdes*, anti-circumvention and copyright infringement are two independent causes of action, then this section says nothing about the interaction between fair use and anti-circumvention. *Reimerdes*, 111 F. Supp. 2d. 294. *Reimerdes* interpreted this silence as the elimination of fair use. *Id.* at 321-22.

[166] 17 U.S.C. § 1201(a)(1)(A) (2000).

circumvention from reaching outside the statutory subject matter of copyright.[167] Second, section 1201(f) limits anti-circumvention indirectly by allowing reverse engineering for only limited purposes.

1. One Difficulty and Danger of the DMCA

One difficulty and danger of the DMCA is that it fails to address what happens when an anti-circumvention technology is used to control access to both proper and improper copyright subject matter. For example, copyright protects movies and video game content,[168] but does not protect the method of operation of the movie player or game console.[169] Two important questions are: (1) what if you cannot separate the content from the platform;[170] and (2) what if analyzing the method of operation of the platform necessarily entails circumventing copyright protection technology? These questions were asked and answered in the video game context by *Atari*, *Accolade*, and *Connectix.* The answer, consistent with section 102(b), was that copyright could not be used to block access to the platform. These cases were decided before the DMCA and the questions must be asked and answered again.

2. The Act of Circumvention and Anti-Circumvention Technology

The anti-circumvention requirements that trigger the DMCA are easily satisfied. The anti-circumvention provisions can be applied when a person "circumvents a technological measure that effectively controls access to a work protected under this title."[171] There are two key ideas: the act (what has to be done to count as circumvention) and the technology (what counts as an anti-circumvention device). The statutory definition of the act of circumvention amounts to bypassing the technological measure without the copyright owner's permission.[172] The technological measure that counts as an anti-circumvention device also

[167] Six years before the DMCA was legislated, the Federal Circuit declared that "[a]n author cannot acquire patent-like protection by putting an idea, process, or method of operation in an unintelligible format and asserting copyright infringement against those who try to understand that idea, process, or method of operation." *Atari I*, 975 F.2d at 837.

[168] 17 U.S.C. § 102(a)(6) (2000).

[169] *Id.* at § 102(b).

[170] That is, what if the media and the player interact in such a way that it is impossible to draw a line between the two?

[171] *Id.* at § 1201(a)(1)(A).

[172] "[T]o 'circumvent a technological measure' means to descramble a scrambled work, to decrypt an encrypted work, or otherwise to avoid, bypass, remove, deactivate, or impair a technological measure, without the authority of the copyright owner." *Id.* at § 1201(a)(3)(A). This is a kind of digital trespass.

amounts to nothing more than the permission of the copyright owner.[173] Both hinge on having or not having permission. Significantly, this means circumvention violations can be brought against anyone who accesses[174] digital work, in a manner that is not in the ordinary course of operation,[175] without the permission of the copyright owner. Thus, circumvention as outlawed by the DMCA is almost indistinguishable from our first definition of reverse engineering.[176]

3. Reverse Engineering Restrictions

Reverse engineering is written into the DMCA as a subsection of, and exception to, the circumvention prohibitions.[177] When copyright can be used to prevent access to functional elements or methods of operation, copyright risks going beyond its statutory subject matter. It is here that reverse engineering provides a necessary "safety valve."[178] However, the text of the DMCA limits and qualifies reverse engineering to the point where it is questionable if it exists as a useful option at all.

The DMCA limits reverse engineering by restricting the act, the means, and the publication of results. The act is limited by who can do the reverse engineering ("a person"), what their purpose must be ("for the sole purpose of . . . interoperability"), how much they may reverse engineer (only the "elements of the program that are necessary"), what kind of devices their results must be directed toward ("an independently created computer program"), what kind of information they are allowed to look for (only information that has "not previously been readily available to the person"), and how to do the work (such that the acts "do not constitute infringement under this title").[179]

The DMCA also restricts the means that can be used (only those "necessary to achieve such interoperability"), and how the means may be used (such that they "do not constitute infringement under this title").[180]

The DMCA further restricts the publication of the reverse engineering results by who can publish ("the person" who did the reverse

[173] "[A] technological measure 'effectively controls access to a work' if the measure, in the ordinary course of its operation, requires the application of information, or a process or a treatment, with the authority of the copyright owner, to gain access to the work." *Id.* at § 1201(a)(3)(B).

[174] The word "access" is this phrase is an attempt at a neutral way of saying view, listen to, play, examine, or use content which had been scrambled, encrypted, or somehow protected.

[175] 17 U.S.C. § 1201(a)(3)(B) (2000).

[176] *See supra* note 1.

[177] Title 17 Section 1201 is entitled "Circumvention of copyright protection systems". Subsection (f) provides "Reverse engineering" exceptions to "infringement under this title". 17 U.S.C. § 1201 (2000).

[178] Universal City Studios v. Reimerdes, 111 F. Supp. 2d at 322 (S.D.N.Y. 2000).

[179] 17 U.S.C. § 1201(f)(1) (2000).

[180] *Id.* at § 1201(f)(2).

engineering), why they can publish ("solely for the purpose of enabling interoperability"), what kind of interoperability the publication must be directed toward ("an independently created computer program"), and how the publication may occur (such that publication does "not constitute infringement under this title or violate applicable law other than this section").[181]

At a minimum, these restrictions turn reverse engineering into a crime if your purpose is anything other than interoperability with an independently created computer program.[182] Even then the restrictions make reverse engineering a dangerous practice.[183] The following case, *Universal City Studios v. Reimerdes*, although not about video games, presents an early interpretation of a few of the reverse engineering provisions of the DMCA.[184]

B. Universal City Studios v. Reimerdes

In late September 1999, Jon Johansen, a 15-year-old Norwegian, and two others reverse engineered a licensed DVD player and discovered the keys to the encryption algorithm.[185] They used this information to create DeCSS,[186] a small program that decrypts DVD movies so they could play them on their own Linux player.[187] In November 1999, the defendants, including Reimerdes, posted DeCSS on the Internet for download and also provided links to other DeCSS-based software.[188] Universal City Studios sued Reimerdes under the DMCA's anti-circumvention sections. This case primarily deals with section 1201(a)(2) of the DMCA, the anti-trafficking provision that "bans offering or providing technology that may be used to circumvent technological means of controlling access to copyrighted works."[189] Reimerdes raised a number of defenses, all of which failed. The defense of fair use, which "has been viewed by courts as a safety valve" to temper copyright rights,[190] failed because fair use is a defense to copyright infringement, and

[181] *Id.* at § 1201(f)(3).

[182] "A 'computer program' is a set of statements or instructions to be used directly or indirectly in a computer in order to bring about a certain result." *Id.* at § 101.

[183] It is not clear at this point, what kind of useful reverse engineering is allowed.

[184] *See* Universal City Studios v. Reimerdes, 111 F. Supp. 2d. 294.

[185] *Id.* at 311. Jon Johanson was recently tried in Norway and cleared of all charges. *See DECSS Author Jon Johansen found Innocent in Norwegian Court*, 2600 NEWS (Jan. 7, 2003), *at* http://www.2600.com/news/view/article/1485. Johansen's lawyer is quoted saying "when you have bought a film legally, you have access to its content, [i]t is irrelevant how you get that access. You have bought the movie after all." *Id.*

[186] DVDs are encrypted with an algorithm called CSS.

[187] Universal City Studios v. Reimerdes, 111 F. Supp. 2d at 311.

[188] *Id.* at 312

[189] *Id.* at 319.

[190] *Id.* at 322.

defendants "are not here sued for copyright infringement."[191] They were sued for anti-circumvention trafficking, not copyright infringement, and the court held that Congress provided no fair use defense to this action.[192]

The defense based on the reverse engineering provisions of the DMCA also failed. Defendants contend that DeCSS is necessary to achieve interoperability and create a DVD player on Linux.[193] The court found that section 1201(f)(3) allows only the individual who did the reverse engineering, i.e. Jon Johansen, to make DeCSS available to others.[194] And even then, the court speculated that Johansen could not post DeCSS because there would be other possible uses for the code, and therefore the posting would not be done "solely to achieve interoperability with Linux or anything else."[195]

Reimerdes shifts power to copyright holders by reading the anti-circumvention provisions of the DMCA to eliminate fair use and eviscerate it's reverse engineering allowances.

C. Video Game Reverse Engineering Under the DMCA

The DMCA will hurt the video game industry by curtailing reverse engineering. The DMCA's anti-circumvention provisions can be easily brought to bear on any digital work because of the ease with which anti-circumvention technology can be wrapped around that work. The DMCA does not require that digital works be well protected—simple scrambling of data is enough. *Reimerdes* found that even a "weak cipher" effectively controls access to copyrighted works.[196] The *Reimerdes* test for what counts as an access control device is circular, and easily met: "if its function is to control access," then it effectively controls access.[197]

Once anti-circumvention technology is added, that work can only be reverse engineered for the narrow purpose of interoperability defined in section 1201(f) of the DMCA. As discussed above, reverse engineering in the video game field is used for more than just

[191] *Id.* at 322.

[192] Universal City Studios v. Reimerdes, 111 F. Supp. 2d at 322.

[193] *Id.* at 320. There was no Linux-based DVD player at the time.

[194] *Id.* at 320.

[195] *Id.* The court reasoned that because DeCSS runs on Windows machines, DeCSS could not be used solely for Linux interoperability. *Id.* This is a dangerous idea because algorithms, source code, and programs can be compiled and run on any number of platforms and by their nature are rarely locked into a single platform.

[196] *Id.* at 317. The court in *RealNetworks, Inc. v. Streambox, Inc.*, found a preliminary "Secret Handshake" qualified as a DMCA anti-circumvention device. 2000 U.S. Dist. LEXIS 1889 (W.D. Wash. 2000).

[197] Universal City Studios v. Reimerdes, 111 F. Supp. 2d at 318.

interoperability—reverse engineering opens up both horizontal and vertical access and enables technological, design, and user interface leap-frogging.[198] This leap-frogging produces fast paced innovation that has created an industry where no single company has remained dominant.[199] Great games drive the industry, and the next great game can come from anywhere. Some games are successful because they push technology forward, other games are successful because they push a design forward. The industry is better because this generation of games improved directly on the previous generation. The legal environment plays an important role in keeping the industry competitive and growing. The DMCA stifles reverse engineering and will slow the industry's growth.

Also troubling is the possible ability of the DMCA to extend protection to un-copyrightable elements. Copyright protection does not extend to the ideas, processes, procedures, and some of the other elements in a work.[200] However, with the addition of simple anti-circumvention technology, the DMCA might be used to block access to unprotected elements much the same as Cohen and Lemley warn patents could be used to block access to "unpatented components."[201]

To add to this trouble, compare older video game hardware[202] to Loffredo's new hardware, discussed above, which uses programmable logic parts.[203] This new technology creates a problem: that which has historically been tangible hardware, visible to the eye, susceptible to probing, experimentation, and included many un-copyrightable elements,[204] is now entirely a software "bitstream"[205] which is easily scrambled and wrapped in anti-circumvention technology. Should hardware be subjected to different copyright treatment because it is programmable? According to section 102 of the Copyright Act, the

[198] Successful game designs and user interfaces quickly become widespread and improved on.

[199] *Changing the game*, THE ECONOMIST, Dec. 6th, 2003, *available at* 2003 WL 58585083 (since its inception, the industry has operated in approximately 5 year cycles with no single manufacturer or game maker able to hold onto the top spot). Conversely, the DMCA could aid industry consolidation.

[200] "In no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work." 17 U.S.C. § 102(b) (2000).

[201] Cohen & Lemley, *supra* note 13, at 26.

[202] Older hardware contained many standard commodity parts such as resistors, transistors, capacitors, and standard integrated circuits.

[203] Loffredo, *supra* note 73.

[204] Although some elements of older hardware have been protected under copyright, "mask works," for example, many standard parts such as resistors, transistors, capacitors, and standard integrated circuit chips have not been subject to copyright protection.

[205] *Id.*

answer is no.[206] Arriving at that result applying the DMCA is sure to be a difficult task.

The DMCA has undone the reverse engineering allowances of *Atari*, *Accolade*, and *Connectix*, where the court carefully balanced copyright holder's rights, fair use, and competition. The DMCA creates a new cause of action, anti-circumvention, that trumps fair use, avoids balancing, and ignores competition. *Atari*, *Accolade*, and *Connectix* articulate a pro-competitive reverse engineering doctrine that allows access to platforms and the ability to create alternative compatible platforms. The DMCA wipes out this balance.

CONCLUSION

Reverse engineering is used by game designers, software and hardware engineers, and is essential to the health of a competitive video game industry. The DMCA's anti-circumvention provisions upset the balance between the rights of copyright holders, fair use, and competition. Because of the importance of reverse engineering, Congress should amend the DMCA to expand allowances for reverse engineering practices.

[206] 17 U.S.C. § 102(b) states that un-copyrightable subject matter can not be made into copyrightable subject matter by changing "the form in which it is described, explained, illustrated, or embodied."

APPENDIX A
UNDERSTANDING CODE

This appendix is presented to illustrate the connection between source code, object code, and disassembled object code.

1) Here is a simple program, written in C. This is source code. Source code is normally not shipped with a product and is often carefully guarded. If run on most computers, this program would print the message "hello, world":

```
/*  This is the "first program" you ever write.  */
/*  Brian W. Kerninghan and Dennis M. Ritchie,*/
/*  The C Programming Language page 5 (2d ed. 1998). */
#include <stdio.h>
main()
{
            printf("hello, world\n");
}
```

2) Here is an example of the object code generated from the above source code. This is what is shipped in games. This is the kind of code (although for a different system) that reverse engineers deciphered in *Sega v. Accolade* and *Sony v. Connectix*. Most consider this, incorrectly, to be 1s and 0s and unreadable. The last line is data that most programmers can read. It contains the words "hello, world." For example, 68 = 'h' 65 = 'e' 6C = 'l' etc.

```
00401010  55  8B  EC 83  EC 40  53  56  57  8D 7D C0  B9  10  00  00  00  B8
00401022  CC CC CC CC F3  AB 68  1C  00  42  00  E8  2E  00  00  00  83  C4
00401034  04  5F  5E  5B  83  C4 40  3B  EC E8  9E  00  00  00  8B  E5  5D C3
0042001C 68  65  6C  6C  6F  2C  20  77  6F  72  6C  64  0A  00  00  00  00  00
```

3) Here is the disassembly of that same object code. The disassembler has taken the object code and reformatted it to improve its readability. All programmers who know x86 assembler can read and understand this code. This is the same program as in 1) and 2).

```
00401010    push   ebp
00401011    mov    ebp,esp
00401013    sub    esp,40h
00401016    push   ebx
00401017    push   esi
00401018    push   edi
00401019    lea    edi,[ebp-40h]
0040101C    mov    ecx,10h
00401021    mov    eax,0CCCCCCCCh
00401026    rep    stosdword ptr [edi]
00401028    push   offset string "hello, world\n" (0042001c)
0040102D    call   printf (00401060)
00401032    add    esp,4
00401035    pop    edi
00401036    pop    esi
00401037    pop    ebx
00401038    add    esp,40h
0040103B    cmp    ebp,esp
0040103D    call   __chkesp (004010e0)
00401042    mov    esp,ebp
00401044    pop    ebp
00401045    ret
```