

31:2 BERKELEY TECHNOLOGY LAW JOURNAL

2016

Pages

1215

to

1514

Berkeley Technology Law Journal

Volume 31, Number 2

Production: Produced by members of the *Berkeley Technology Law Journal*.
All editing and layout done using Microsoft Word.

Printer: Joe Christensen, Inc., Lincoln, Nebraska.

Printed in the U.S.A.

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Library Materials, ANSI Z39.48—1984.

Copyright © 2016 Regents of the University of California.

All Rights Reserved.

Berkeley Technology Law Journal

University of California

School of Law

3 Boalt Hall

Berkeley, California 94720-7200

btlj@law.berkeley.edu

<http://www.btlj.org>



BERKELEY TECHNOLOGY LAW JOURNAL

VOLUME 31

NUMBER 2

2016

TABLE OF CONTENTS

ARTICLES

FUNCTIONALITY AND EXPRESSION IN COMPUTER PROGRAMS: REFINING THE TESTS FOR SOFTWARE COPYRIGHT INFRINGEMENT	1215
<i>Pamela Samuelson</i>	
PATENT NATIONALLY, INNOVATE LOCALLY	1301
<i>Camilla A. Hrdy</i>	
USING ANTITRUST LAW TO CHALLENGE TURING’S DARAPRIM PRICE INCREASE.....	1379
<i>Michael A. Carrier, Nicole L. Levidow, Aaron S. Kesselheim</i>	
WIRELESS NETWORK NEUTRALITY: TECHNOLOGICAL CHALLENGES AND POLICY IMPLICATIONS.....	1409
<i>Christopher S. Yoo</i>	
COPYRIGHT REFORM AND COPYRIGHT MARKET: A CROSS-PACIFIC PERSPECTIVE	1461
<i>Jiarui Liu</i>	

SUBSCRIBER INFORMATION

The *Berkeley Technology Law Journal* (ISSN1086-3818), a continuation of the *High Technology Law Journal* effective Volume 11, is edited by the students of the University of California, Berkeley, School of Law (Boalt Hall) and is published in print three times each year (March, September, December), with a fourth issue published online only (July), by the Regents of the University of California, Berkeley. Periodicals Postage Rate Paid at Berkeley, CA 94704-9998, and at additional mailing offices. POSTMASTER: Send address changes to Journal Publications, University of California, Berkeley Law—Library, LL123 Boalt Hall—South Addition, Berkeley, CA 94720-7210.

Correspondence. Address all correspondence regarding subscriptions, address changes, claims for non-receipt, single copies, advertising, and permission to reprint to Journal Publications, University of California, Berkeley Law—Library, LL123 Boalt Hall—South Addition, Berkeley, CA 94705-7210; (510) 643-6600; Journals@law.berkeley.edu. *Authors:* see section titled Information for Authors.

Subscriptions. Annual subscriptions are \$65.00 for individuals and \$85.00 for organizations. Single issues are \$30.00. Please allow two months for receipt of the first issue. Payment may be made by check, international money order, or credit card (MasterCard/Visa). Domestic claims for non-receipt of issues should be made within 90 days of the month of publication; overseas claims should be made within 180 days. Thereafter, the regular back issue rate (\$30.00) will be charged for replacement. Overseas delivery is not guaranteed.

Form. The text and citations in the *Journal* conform generally to the THE CHICAGO MANUAL OF STYLE (16th ed. 2010) and to THE BLUEBOOK: A UNIFORM SYSTEM OF CITATION (Columbia Law Review Ass’n et al. eds., 20th ed. 2015). Please cite this issue of the *Berkeley Technology Law Journal* as 31 BERKELEY TECH. L.J. ____ (2016).

BTLJ ONLINE

The full text and abstracts of many previously published *Berkeley Technology Law Journal* articles can be found at <http://www.btlj.org>. Our site also contains a cumulative index; general information about the *Journal*; the *Bolt*, a collection of short comments and updates about new developments in law and technology written by BTLJ members; and *BTLJ Commentaries*, an exclusively online publication for pieces that are especially time-sensitive and shorter than typical law review articles.

INFORMATION FOR AUTHORS

The Editorial Board of the *Berkeley Technology Law Journal* invites the submission of unsolicited manuscripts. Submissions may include previously unpublished articles, essays, book reviews, case notes, or comments concerning any aspect of the relationship between technology and the law. If any portion of a manuscript has been previously published, the author should so indicate.

Format. Submissions are accepted in electronic format through the ExpressO online submission system. Authors should include a curriculum vitae and resume when submitting articles, including his or her full name, credentials, degrees earned, academic or professional affiliations, and citations to all previously published legal articles. The ExpressO submission website can be found at <http://law.bepress.com/expresso>.

Citations. All citations should conform to THE BLUEBOOK: A UNIFORM SYSTEM OF CITATION (Columbia Law Review Ass'n et al. eds., 20th ed. 2015).

Copyrighted Material. If a manuscript contains any copyrighted table, chart, graph, illustration, photograph, or more than eight lines of text, the author must obtain written permission from the copyright holder for use of the material.

DONORS

The *Berkeley Technology Law Journal* and the Berkeley Center for Law & Technology acknowledge the following generous donors to Berkeley Law's Law and Technology Program:

Partners

COOLEY LLP

FENWICK & WEST LLP

ORRICK, HERRINGTON &
SUTCLIFFE LLP

Benefactors

COVINGTON & BURLING LLP

SIDLEY AUSTIN LLP

FISH & RICHARDSON P.C.

SKADDEN, ARPS, SLATE, MEAGHER
& FLOM LLP & AFFILIATES

KASOWITZ BENSON
TORRES & FRIEDMAN LLP

VAN PELT, YI & JAMES LLP

KIRKLAND & ELLIS LLP

WEIL, GOTSHAL & MANGES LLP

LATHAM & WATKINS LLP

WHITE & CASE LLP

MCDERMOTT WILL & EMERY

WILMER CUTLER PICKERING HALE
AND DORR LLP

MORRISON & FOERSTER LLP

WILSON SONSINI
GOODRICH & ROSATI

WINSTON & STRAWN LLP

Members

BAKER BOTTS LLP

KILPATRICK TOWNSEND &
STOCKTON LLP

BAKER & MCKENZIE LLP

KNOBBE MARTENS
OLSON & BEAR LLP

DURIE TANGRI LLP

LEE TRAN & LIANG LLP

GJEL ACCIDENT ATTORNEYS

MUNGER, TOLLES & OLSON LLP

GTC LAW GROUP LLP & AFFILIATES

O'MELVENY & MYERS LLP

GUNDERSON DETTMER STOUGH
VILLENEUVE FRANKLIN &
HACHIGIAN, LLP

PAUL HASTINGS LLP

HAYNES AND BOONE, LLP

ROPES & GRAY LLP

HOGAN LOVELLS LLP

SIMPSON THACHER & BARTLETT LLP

IRELL & MANELLA LLP

TURNER BOYD LLP

KEKER & VAN NEST LLP

WEAVER AUSTIN VILLENEUVE &
SAMPSON, LLP

BOARD OF EDITORS

2015–2016

Executive Committee

Editor-in-Chief

JOSHUA D. FURMAN

Managing Editor

MISHA TSUKERMAN

Senior Articles Editors

RAVI ANTANI
ZACHARY FLOOD
KELLY VARGAS

Senior Executive Editor

DIANA OBRADOVICH

Senior Scholarship Editor

CAROLINA GARCIA

Senior Annual Review Editors

GINNY SCHOLTES
SORIN ZAHARIA

Senior Online Content Editor

NOAH DRAKE

Editorial Board

Commentaries Editors

SWAROOP POUDEL
YU TANEBE
BONNIE WATSON

Production Editors

ROXANA GUIDERO
DUSTIN VANDENBERG

Technical Editors

KRISTINA PHAM
CHRISTOPHER YANDEL

Annual Review Editors

CYNTHIA LEE
BENJAMIN LI

Notes & Comments Editors

WAQAS AKMAL
ERICA FISHER

Symposium Editors

TOMMY BARCZYK
JOHN RUSSELL

Submissions Editors

PHILIP MERKSAMER
MAYA ZIV

Web & Technology Editors

JOE CRAIG
LEIGHANNA MIXTER

External Relations Editor

SIMONE FRIEDLANDER

Member Relations Editor

STEPHANIE CHENG

Alumni Relations Editor

GOLDA CALONGE

Web Content Editor

FAYE WHISTON

JESSICA ANNIS
JOEL BROUSSARD
CHRISTIAN CHESSMAN
DANEILLE DEVLIN
KELSEY GUANCIALE

Articles Editors

YESOL HAN
CASSY HAVENS
MARK JOSEPH
BILAL MALIK

CHRIS NORTON
LIDA RAMSEY
ERIC RIEDEL
ARIEL ROGERS
MAX SLADEK DE LA CAL

MEMBERSHIP

Vol. 31 No. 2

Associate Editors

WILL BINKLEY	PAMUDH KARIYAWASAM	ALEJANDRO ROTHAMEL
BRITTANY BRUNS	HILARY KRASE	LUKAS SIMAS
DAPHNE CHEN	JOYCE LI	RICHARD SIMS
YUHAN (ALICE) CHI	MATTHEW MCCLELLAN	FERNANDA SOLIS CAMARA
KEVIN CHIU	GAVIN MOLER	SARAH SUWANDA
SARA CHUGH	CATALINA MONCADA	JON TANAKA
NOA DREYMANN	SHELBY NACINO	VALERIE TRUONG
RAN DUAN	JUAN NAZAR	RAOUL GRIFONI- WATERMAN
JORDAN FRABONI	ROBERT OLSEN	REID WHITAKER
JEREMY ISARD	EUNICE PARK	
NATASA JANEZ	JAIDEEP REDDY	

Members

YASMINE AGELIDIS	NOAH GUINEY	MICHELLE PARK
NIKI BAWA	BRIAN HALL	DYLAN PETERSON
KATE BRIDGE	ANDREA HALL	KEVIN PORMIR
JESSICA BRODSKY	JESSICA HOLLIS	CHRISTELLE PRIDE
KELSEA CARLSON	JENNIFER HSU	JING XUN QUEK
STEPHEN CHAO	KENSUKE INOUE	LEE REDFEARN
JOSEPH CHRISTIE	HYE JIN KIM	MATT RICE
RACHEL CORRIGAN	RITHIKA KULATHILA	FAITH SHAPIRO
MICHAEL DEAMER	SARAH KWON	DARINA SHTRAKHMAN
THOMAS DEC	TIFFANY LEUNG	JOSHUA STEELE
DARIUS DEHGHAN	MEI LIU	EVE TAI
ELENA FALLOON	STASHA LOEZA	MY THAN
MEGHAN FENZEL	SARAH MULLINS	CASEY TONG
EVAN FERGUSON	ELI NESS	ELISSA WALTER
WHITNEY FLORIAN	NATE NGEREBARA	MELISSA WEE
ETHAN FRIEDMAN	PEGGY NI	TAMARA WIESEBRON
ANDREY GAVRILENKO	JESSICA OGLESBEE	SHONG YIN
KAN GU	BARCLAY OUDERSLUYS	YU ZHAO
	ROBERT PARIS	

BTLJ ADVISORY BOARD

JIM DEMPSEY

*Executive Director of the
Berkeley Center for Law & Technology
U.C. Berkeley School of Law*

ROBERT C. BERRING, JR.

*Walter Perry Johnson Professor of Law
U.C. Berkeley School of Law*

MATTHEW D. POWERS

Tensegrity Law Group, LLP

JESSE H. CHOPER

*Earl Warren Professor of Public Law
U.C. Berkeley School of Law*

PAMELA SAMUELSON

*Professor of Law & Information
and Faculty Director of the
Berkeley Center for Law & Technology
U.C. Berkeley School of Law*

PETER S. MENELL

*Professor of Law and Faculty
Director of the Berkeley Center
for Law & Technology
U.C. Berkeley School of Law*

LIONEL S. SOBEL

*Visiting Professor of Law
U.C.L.A. School of Law*

ROBERT P. MERGES

*Wilson Sonsini Goodrich & Rosati
Professor of Law and Faculty
Director of the Berkeley Center
for Law & Technology
U.C. Berkeley School of Law*

LARRY W. SONSINI

Wilson Sonsini Goodrich & Rosati

REGIS MCKENNA

*Chairman and CEO
Regis McKenna, Inc.*

MICHAEL STERN

Cooley LLP

DEIRDRE K. MULLIGAN

*Assistant Professor and Faculty Director
of the Berkeley Center for
Law and Technology
U.C. Berkeley School of Information*

MICHAEL TRAYNOR

Cobalt LLP

JAMES POOLEY

*Deputy Director General of the
World Intellectual Property Organization*

THOMAS F. VILLENEUVE

*Gunderson Dettmer Stough Villeneuve
Franklin & Hachigian LLP*

BERKELEY CENTER FOR LAW & TECHNOLOGY 2015–2016

Executive Director

JIM DEMPSEY

Faculty Directors

KENNETH A. BAMBERGER	PETER S. MENELL	PAMELA SAMUELSON
CATHERINE CRUMP	ROBERT P. MERGES	PAUL SCHWARTZ
CHRIS HOOFNAGLE	DEIRDRE MULLIGAN	JENNIFER URBAN
SONIA KATYAL	MOLLY S. VAN HOUWELING	

Staff Directors

LOUISE LEE	RICHARD FISK	CLAIRE TRIAS
------------	--------------	--------------

FUNCTIONALITY AND EXPRESSION IN COMPUTER PROGRAMS: REFINING THE TESTS FOR SOFTWARE COPYRIGHT INFRINGEMENT

Pamela Samuelson[†]

ABSTRACT

Courts have struggled for decades to develop a test for judging infringement claims in software copyright cases that distinguishes between program expression that copyright law protects and program functionality for which copyright protection is unavailable. The case law thus far has adopted four main approaches to judging copyright infringement claims in software cases. One, now mostly discredited, test would treat all structure, sequence, and organization (SSO) of programs as protectable expression unless there is only one way to perform a program function. A second, now widely applied, three-step test calls for creating a hierarchy of abstractions for an allegedly infringed program, filtering unprotectable elements, and comparing the protectable expression of the allegedly infringed program with the expression in the second program that is the basis of the infringement claim. A third approach has focused on whether the allegedly infringing elements are program processes or methods of operation that lie outside the scope of protection available from copyright law. A fourth approach has concentrated on whether the allegedly infringing elements of a program are instances in which ideas or functions have merged with program expression. This Article offers both praise and criticism of the approaches taken thus far to judging software copyright infringement and proposes an alternative unified test for infringement that is consistent with traditional principles of copyright law and that will promote healthy competition and ongoing innovation in the software industry.

DOI: <https://dx.doi.org/10.15779/Z38WW76Z83>

© 2016 Pamela Samuelson.

[†] Richard M. Sherman Distinguished Professor of Law, Berkeley Law School. I am very grateful to Kathryn Hashimoto for excellent research for and editing of this article. I am also grateful to Clark Asay, Jonathan Band, Joshua Bloch, Oren Bracha, Dan Burk, Julie Cohen, Joe Craig, Charles Duan, Shubha Ghosh, Ariel Katz, Peter Lee, Mark Lemley, Glynn Lunney, Corynne McSherry, Christina Mulligan, Aaron Perzanowski, Michael Risch, Christopher Jon Sprigman, Fred von Lohmann, and Phil Weiser for comments on an earlier draft of this article. I also wish to thank Lionel Bently for the opportunity to give the 10th Annual International IP Lecture at Emanuel College at Cambridge University on which this Article was initially based.

TABLE OF CONTENTS

I.	INTRODUCTION	1217
II.	THE ABSTRACTION-FILTRATION-COMPARISON TEST	1224
A.	THE ROCKY ROAD TO <i>ALTAI</i>	1225
B.	THE <i>ALTAI</i> DECISION AND THE AFC TEST	1230
C.	A CLOSER LOOK AT <i>ALTAI</i> 'S FILTRATION FACTORS	1232
III.	CONCEPTUALIZING THE PROPER ROLE OF § 102(b) IN COMPUTER PROGRAM COPYRIGHT CASES	1237
A.	FIVE UNCONTROVERSIAL PROPOSITIONS ABOUT § 102(b)	1238
1.	<i>Section 102(b) Does Not Exclude Program Code from Protection</i>	1238
2.	<i>The Procedure, Process, System and Method of Operation Exclusions of § 102(b) Must Mean Something</i>	1239
3.	<i>The Process and System Exclusions of § 102(b) Are Partly Aimed at Maintaining Boundaries between Copyright and Patent Laws</i>	1241
4.	<i>Because of § 102(b) Exclusions, the Scope of Copyright in Programs Is Thinner than the Scope of Copyright in Conventional Literary Works</i>	1243
5.	<i>SSO Obscures the Distinction between Nonliteral Elements of Programs That Are Protectable by Copyright and Those That Are Unprotectable Under § 102(b)</i>	1244
6.	<i>Summary</i>	1245
B.	<i>LOTUS V. BORLAND</i>	1245
C.	<i>ORACLE V. GOOGLE</i>	1252
D.	THE IMPLICATIONS OF § 102(b) FOR COMPATIBILITY DEFENSES	1258
IV.	FUNCTIONALITY AND EXPRESSION SOMETIMES MERGE IN SOFTWARE CASES	1267
A.	ORIGINS OF THE MERGER DOCTRINE	1268
B.	THE ROLE OF THE MERGER DOCTRINE IN ARCHITECTURAL WORK AND SOFTWARE CASES	1270
C.	MERGER MAY BE FOUND WHEN A PLAINTIFF'S DESIGN CHOICES SERVE AS CONSTRAINTS ON THE CHOICES AVAILABLE TO SECOND COMERS.....	1275
D.	SOMETIMES PROGRAM FUNCTIONS MERGE WITH PROGRAM CODE.....	1278

E.	THE CAFC ERRED IN INTERPRETING THE MERGER DOCTRINE	1280
V.	DIFFERENT CONCEPTUALIZATIONS ON THE RELATIONSHIP BETWEEN PATENT AND COPYRIGHT PROTECTIONS FOR SOFTWARE	1284
A.	REASONS TO BE CAUTIOUS OF CATEGORICAL EXCLUSIVITY ARGUMENTS ABOUT PATENT AND COPYRIGHT PROTECTIONS FOR SOFTWARE INNOVATIONS.....	1286
B.	THE <i>ORACLE</i> DECISION’S ANALYSIS OF COPYRIGHT- PATENT BOUNDARIES WAS FLAWED	1289
C.	AN ALTERNATIVE APPROACH TO CONCEPTUALIZING THE ROLES OF COPYRIGHTS AND PATENTS IN PROTECTING SOFTWARE INNOVATIONS.....	1291
D.	SOFTWARE DEVELOPERS ATTAIN COMPETITIVE ADVANTAGE BEYOND IP RIGHTS.....	1293
VI.	REFINING THE TEST FOR SOFTWARE COPYRIGHT INFRINGEMENT.....	1294

I. INTRODUCTION

The paradigmatic roles of copyright and patent laws have been, respectively, to protect original authorial expressions from illicit copying, and to protect novel and nonobvious functional designs (if they have been appropriately claimed and examined by patent officials) from illicit uses.¹ It would be convenient if copyright law could be assigned the role of protecting the expression in computer programs and patent law the role of protecting program functionality. While courts continue to try to distinguish between program expression and program functionality, this distinction has proven elusive in the decades since Congress decided to extend copyright protection to computer programs.²

1. See J.H. Reichman, *Legal Hybrids Between the Patent and Copyright Paradigms*, 94 COLUM. L. REV. 2432, 2448–53 (1994) (describing the classical patent and copyright paradigms in the international intellectual property system).

2. While some claim that Congress extended copyright protection to computer programs when it enacted the Copyright Act of 1976, there is some ambiguity in the legislative history on this point. Compare FINAL REPORT OF THE NAT’L COMM’N ON NEW TECHNOLOGICAL USES OF COPYRIGHTED WORKS 15–16 (1978) [hereinafter CONTU REPORT] (concluding that Congress had extended copyright protection to software in 1976) with Pamela Samuelson, *CONTU Revisited: The Case Against Copyright Protection for Computer Programs in Machine-Readable Form*, 1984 DUKE L.J. 663, 694–96 (1984) (suggesting that § 117 in the 1976 Act preserved the status quo of unprotectability under

In the years preceding the enactment of the Copyright Act of 1976,³ members of Congress were warned that the functionality of computer programs would make it difficult to fit them into the copyright realm.⁴ However, lingering concerns about the potential misfit were for a time allayed by a 1978 National Commission on New Technological Uses of Copyrighted Works (CONTU) report that endorsed copyright protection for programs.⁵ CONTU observed that “the distinction between copyrightable computer programs and uncopyrightable processes or methods of operation does not always seem to ‘shimmer with clarity,’” but it was nevertheless “important that the distinction between programs and processes be made clear.”⁶ The report expressed optimism that traditional principles of copyright law, when applied to programs, would strike the right balance,⁷ and it was content to leave the difficult (and perhaps “futile”) task of

the prior act as to computer-related subject matters). However, any such ambiguity was resolved by 1980 amendments to the Copyright Act of 1976 (1976 Act), which implemented legislative changes that CONTU recommended in its report. *See* Pub. L. No. 96-517, 94 Stat. 3015 (codified at 17 U.S.C. §§ 101, 117 (1980)).

3. Pub. L. No. 94-553, 90 Stat. 2541 (1976), codified at 17 U.S.C. § 101, et seq.

4. *See* Hearings Before the Subcomm. on Patents, Trademarks, and Copyrights of the S. Comm. on the Judiciary Pursuant to S. Res. 37 on S. 597, 90th Cong. 192–97 (1967), reprinted in 9 OMNIBUS COPYRIGHT REVISION LEGISLATIVE HISTORY 192–97 (George S. Grossman ed., 1976) (testimony of Professor Arthur Miller). Miller expressed concern that courts might construe copyright protection for programs as “extend[ing] to or embody[ing] the process, scheme, or plan that the program uses to achieve a functional goal,” saying this would confer “patent like protection under the guise of copyright.” *Id.* at 197. Congress responded to these concerns by adopting a provision stating that “[i]n no case does copyright protection for an original work of authorship extend to any . . . procedure, process, system [or] method of operation . . . regardless of the form in which it is . . . embodied in such work.” 17 U.S.C. § 102(b) (2012). This provision is discussed at length *infra* Part III.

5. CONTU REPORT, *supra* note 2, at 1–2. CONTU acknowledged that there was not “universal agreement” about copyright protection for software. *Id.* at 20–21. *See also* Stephen Breyer, *The Uneasy Case for Copyright: A Study of Copyright in Books, Photocopies, and Computer Programs*, 84 HARV. L. REV. 281, 344–46 (1970) (questioning the economic case for extending copyright protection to computer programs). While CONTU was deliberating about new technology issues, the World Intellectual Property Organization was considering a sui generis form of intellectual property protection for software. *See* WORLD INTELLECTUAL PROP. ORG., INT’L BUREAU, MODEL PROVISIONS ON THE PROTECTION OF COMPUTER SOFTWARE (1978). Whether computer programs should be protected by copyright law was not part of CONTU’s original charter, which mainly focused on photocopying and digitizing published texts. *See* Samuelson, *CONTU Revisited*, *supra* note 2, at 663 n.2, 699. That may explain why none of the Commissioners had any expertise about computers or computer programs. *Id.* at 699.

6. CONTU REPORT, *supra* note 2, at 18.

7. *See id.* at 12–23. CONTU thought copyright should grant no more economic power than was needed to create proper incentives to create software. *Id.* at 12.

drawing boundaries between program expression and functionality to the judiciary.⁸ Unfortunately, CONTU failed to fully understand the intrinsic functionality of computer programs, the importance of standards and network effects in the software industry, and the inherent need to develop software capable of interoperating with other programs. It also failed to offer guidance on how, when, and why functionality should constrain the scope of copyright protection in programs.⁹

Commentators have debated for decades how much legal protection software developers should get from copyright law in order to induce optimal levels of investment in the development of computer programs.¹⁰ Some have worried that copyright protection for programs might either be too “weak” if infringement could be easily avoided by rewriting the same

8. *Id.* at 22–23.

9. According to CONTU, programs were no more functional than sound recordings, *id.* at 10, which was simply not true. After all, the inherent purpose of computer programs is to automate functional processes, whereas the purpose of sound recordings is to allow users to listen to music. *See id.* at 27–29 (Hersey dissent distinguishing program functionality from other copyrighted works). CONTU also asserted that utility had never been a bar to the copyrightability of a work or a limit on the scope of protection available to protected works, *id.* at 19–21, which was also untrue. *See* Samuelson, *CONTU Revisited*, *supra* note 2, at 732–39 (explaining reasons why utilitarian works have conventionally been excluded from copyright protection); *see also* Peter S. Menell, *Tailoring Legal Protection for Computer Software*, 39 STAN. L. REV. 1329, 1359–61 (1987) (offering other reasons why copyright protection is inappropriate for operating system software).

10. From the late 1980s to the late 1990s, scholars produced an extensive literature about copyright protection for computer programs, particularly their nonliteral elements. *See, e.g.*, Jane C. Ginsburg, *Four Reasons and a Paradox: The Manifest Superiority of Copyright over Sui Generis Protection of Computer Software*, 94 COLUM. L. REV. 2559 (1994); Dennis S. Karjala, *Copyright, Computer Software, and the New Protectionism*, 28 JURIMETRICS J. 33 (1987); Peter S. Menell, *An Analysis of the Scope of Copyright Protection for Application Programs*, 41 STAN. L. REV. 1045 (1989); Arthur R. Miller, *Copyright Protection for Computer Programs, Databases, and Computer-Generated Works: Is Anything New Since CONTU?*, 106 HARV. L. REV. 977 (1993); J.H. Reichman, *Computer Programs as Applied Scientific Know-How: Implications of Copyright Protection for Commercialized University Research*, 42 VAND. L. REV. 639 (1989); Pamela Samuelson et al., *A Manifesto Concerning the Legal Protection of Computer Programs*, 94 COLUM. L. REV. 2308 (1994); Lloyd L. Weinreb, *Copyright for Functional Expression*, 111 HARV. L. REV. 1150 (1998); Steven R. Englund, Note, *Idea, Process, or Protected Expression?: Determining the Scope of Copyright Protection of the Structure of Computer Programs*, 88 MICH. L. REV. 866 (1990). Especially influential was a law review article, later incorporated into the Nimmer treatise. *See* David Nimmer et al., *A Structured Approach to Analyzing Substantial Similarity of Computer Software in Copyright Infringement Cases*, 20 ARIZ. ST. L.J. 625 (1988); MELVILLE B. NIMMER & DAVID NIMMER, NIMMER ON COPYRIGHT § 13.03 [F](2015) [hereinafter NIMMER ON COPYRIGHT]; *see also* JONATHAN BAND & MASANOBU KATOH, INTERFACES ON TRIAL (1995); JONATHAN BAND & MASANOBU KATOH, INTERFACES ON TRIAL 2.0 (2011).

program in different source code, or too “strong” if programmers felt compelled to do things differently than an existing program in order to avoid infringement, thereby impeding beneficial standardization.¹¹ That concern has manifested itself in the software copyright cases that followed.

Appellate courts have taken four main approaches to distinguishing the original expression in computer programs from program functionality. A first-in-time, but now much discredited, approach was adopted by the Third Circuit Court of Appeals in *Whelan Associates v. Jaslow Dental Lab., Inc.*, under which the “structure, sequence, and organization” (SSO) of computer programs was deemed protectable expression unless there was only one way to perform a function (in which case a second comer could use the same SSO under the merger of idea and expression doctrine).¹²

A second was the Second Circuit Court of Appeals’ approach in *Computer Associates Int’l, Inc. v. Altai, Inc.*¹³ The *Altai* decision was highly critical of *Whelan* and its test for software copyright infringement.¹⁴ As an alternative, *Altai* offered the abstraction-filtration-comparison (AFC) test for judging copyright infringement claims in computer program cases.¹⁵ *Altai*’s principal contribution has been its insistence that courts must “filter” out unprotectable elements of programs, such as those necessary for achieving interoperability with other programs, before assessing infringement claims.¹⁶ The AFC test has been adopted and applied in numerous subsequent cases.¹⁷

11. See, e.g., Breyer, *supra* note 5, at 347–48; see also Samuelson et al., *Manifesto*, *supra* note 10, at 2356–63 (explaining why applying copyright law to computer programs might lead to cycles of under- and over-protection).

12. 797 F.2d 1222, 1236, 1248 (3d Cir. 1986), *cert. denied*, 479 U.S. 1031 (1987). *Whelan* is discussed *infra* text accompanying notes 53–70.

13. 982 F.2d 693 (2d Cir. 1992).

14. *Altai*’s criticism of *Whelan* is discussed *infra* text accompanying notes 82–85.

15. *Altai*, 982 F.2d at 706–11.

16. *Id.* at 707–10.

17. The *Altai* approach to software copyright infringement has been endorsed in numerous other circuit court decisions. See, e.g., *Sega Enters. Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1524–25 (9th Cir. 1992); *Gates Rubber Co. v. Bando Chem. Indus.*, 9 F.3d 823, 834 (10th Cir. 1993); *Bateman v. Mnemonics, Inc.*, 79 F.3d 1532, 1543–44 (11th Cir. 1996). As of July 6, 2015, *Altai* had been positively cited in 389 federal court decisions on the subject of copyright in the Lexis database and had been cited in all circuits. See also Mark A. Lemley, *Convergence in the Law of Software Copyright?*, 10 BERKELEY TECH. L.J. 1, 14–15 (1995) (treating *Altai* as the leading case on copyright protection for computer programs).

A third approach was used by the First Circuit Court of Appeals in *Lotus Dev. Corp. v. Borland Int'l, Inc.*¹⁸ *Borland* ruled that the command hierarchy of a spreadsheet program was an integral part of a method of operation that 17 U.S.C. § 102(b) excluded from the scope of copyright protection in programs.¹⁹ The court compared the Lotus command structure to the clearly uncopyrightable set of buttons that control the operation of videotape machines.²⁰

A fourth approach emerged in the Sixth Circuit Court of Appeals' decision in *Lexmark Int'l, Inc. v. Static Control Components*.²¹ The court decided that a computer program embedded in Lexmark's printer cartridges that competitors had to install to enable cartridges to interoperate with Lexmark printers was ineligible for copyright protection.²² Because the idea or function of the program and its expression had merged, the program was held to be uncopyrightable.²³

A common thread through the *Altai*, *Borland*, and *Lexmark* decisions is that copyright infringement does not occur when a second comer must copy some aspects of another firm's program to achieve compatibility. Courts have deemed the functional requirements for achieving compatibility to be unprotectable elements of these copyrighted programs, even though more than a modicum of creativity may have imparted originality to these elements.

The seeming consensus that program interfaces necessary for interoperability are unprotectable by copyright law was recently called into question by the Court of Appeals for the Federal Circuit (CAFC) in *Oracle Am., Inc. v. Google Inc.*²⁴ At issue was whether the command structure of certain elements of the Java application program interface (API) was protectable by copyright law. The CAFC reversed a lower court ruling that this command structure was an unprotectable method of operation, or

18. 49 F.3d 807, 814–15 (1st Cir. 1995), *aff'd by an equally divided Court*, 516 U.S. 233 (1996)

19. *Id.* at 815–18. That section provides that “[i]n no case does copyright protection for an original work of authorship extend to any . . . procedure, process, system [or] method of operation . . . regardless of the form in which it is . . . embodied in such work.” 17 U.S.C. § 102(b). The *Borland* decision is discussed at length in Section III.B.

20. *Borland*, 49 F.3d at 817.

21. 387 F.3d 522 (6th Cir. 2004).

22. *Id.* at 542–43.

23. *Id.* at 541–42. The *Lexmark* decision is discussed in Section IV.D.

24. 750 F.3d 1339 (Fed. Cir. 2014), *cert. denied*, 135 S. Ct. 2887 (2015). The CAFC's *Oracle* decision is discussed at length in Sections III.C & D, Section IV.E, and Section V.B.

alternatively that copyright protection was unavailable under the merger doctrine.²⁵

Although the CAFC in *Oracle* purported to defer to *Altai*,²⁶ the ghost of the discredited *Whelan* decision reappeared in the CAFC's endorsement of the copyrightability of program SSO.²⁷ Like the Third Circuit in *Whelan*, the CAFC in *Oracle* viewed § 102(b) as merely a restatement of the idea/expression distinction and treated the merger doctrine as applicable only when there was, *ex ante*, no other way for engineers to design this or other program SSO.²⁸ The CAFC was untroubled by the prospect that software developers might obtain both patent and copyright protection for APIs of computer programs.²⁹ There was, in its view, no need to sort out functionality and expression in computer programs. Copyright could protect both as long as there was a modicum of creativity to support the claim of copyright.

The *Oracle* decision has rekindled a decades-old debate, which many thought had been settled in the late 1990s, about the proper scope of copyright protection for computer programs and how courts should analyze claims of software copyright infringement.³⁰ The Supreme Court decision not to review *Oracle* leaves the CAFC ruling intact for the time being.³¹

25. *Oracle Am., Inc. v. Google Inc.*, 872 F. Supp. 2d 974, 977, 998–1002 (N.D. Cal. 2012), *rev'd*, 750 F.3d 1339 (Fed. Cir. 2014).

26. *Oracle*, 750 F.3d at 1355–58.

27. *See id.* at 1366–68.

28. *See id.* at 1359–62 (discussing the merger doctrine), 1364–68 (discussing § 102(b)).

29. *See id.* at 1380–81.

30. The *Oracle* decision has already spawned new rounds of software copyright litigation. *See, e.g.*, *Cisco Sys., Inc. v. Arista Networks, Inc.*, No. 5:14-cv-05344 (N.D. Cal. 2014); *Synopsys, Inc. v. ATopTech, Inc.*, No. 3:13-cv-02965 (N.D. Cal. 2013). The complaints in both cases include patent as well as copyright infringement claims. *See* Complaint for Copyright and Patent Infringement, *Cisco Sys., Inc. v. Arista Networks, Inc.*, No. 14-5344, 2014 WL 6844640 (N.D. Cal. Dec. 5, 2014); Amended Complaint for Copyright Infringement, Patent Infringement, Breach of Contract, and Breach of Implied Covenant of Good Faith and Fair Dealing, *Synopsys, Inc. v. ATopTech, Inc.*, No. 3:13-cv-02965-MMC, 2013 WL 7117632 (N.D. Cal. Nov. 25, 2013). Appeals from the District Court decisions in *Arista* and *Synopsys* will go to the CAFC because the plaintiffs in those cases have learned a lesson from the *Oracle* case that tacking on a patent claim will avoid going to the Ninth Circuit where compatibility and § 102(b) defenses would be more likely to prevail.

31. 750 F.3d 1339 (Fed. Cir. 2014), *cert. denied*, 135 S. Ct. 2887 (2015). The CAFC's ruling will not be binding on the Ninth Circuit or other courts. The *Oracle* case was remanded for retrial of Google's fair use defense, resulting in a jury verdict in favor of Google and fair use. *Oracle Am., Inc. v. Google Inc.*, No. C 10-03561 WHA, 2016 WL 3181206 (N.D. Cal. June 8, 2016), *appeal docketed*, Nos. 17-1118, -1202 (Fed. Cir. Nov. 14, 2016).

This Article aims to provide guidance about how courts should assess claims of copyright infringement in computer program cases. It assesses the strengths and limitations of the various tests for infringement adopted in software copyright cases and offers a refined test for infringement that takes the soundest features from the existing tests and consolidates them into one unified approach.

Part II reviews the *Whelan* and *Altai* decisions and explains why the AFC test is more consistent with traditional principles of copyright law than the *Whelan* is-there-any-other-way-to-do-it test. *Altai* recognized that external factors, such as the need to be compatible with other programs, sometimes constrain the design decisions of subsequent programmers, and when this happens, those constraints limit the scope of copyright protection in programs. While there is much in the decision to praise, *Altai* failed to heed the statutory directive in § 102(b) that procedures, processes, systems, and methods of operation should also be filtered out before making judgments on copyright infringement claims in software cases.

Part III explains the important role that § 102(b) has played in various computer program cases, including *Borland*. It then discusses the numerous respects in which the CAFC in *Oracle* misinterpreted § 102(b). It considers six types of cases in which courts have held that aspects of programs that are necessary for achieving interoperability with other programs or hardware are too functional to be protected by copyrights.

Part IV explains why the merger doctrine has an important role to play in the assessment of infringement claims involving computer programs and why the CAFC erred in its interpretation of this doctrine. Courts should explicitly recognize a merger of function and expression doctrine in some computer program cases. That doctrine would usefully complement an analysis of elements that may be unprotectable under § 102(b).

Part V considers the roles that copyright and patent law should play in protecting program innovations, with particular attention to how courts should assess claims that copyright protection should be unavailable to aspects of programs possibly eligible for patent protection. The CAFC in *Oracle* conflated copyright and utility patent protections for software, as though it was unnecessary to even try to distinguish program expression and functionality.

Part VI offers a pragmatic approach to distinguishing between program functionality and expression in copyright cases, as well as a refined version of the *Altai* AFC test that is consistent with the traditional principles of copyright law and the overwhelming majority of software copyright cases (even if not consistent with the CAFC's *Oracle* decision). Competition and ongoing innovation will better thrive when the scope of copyright protection

is relatively thin, allowing programmers to reuse functional design elements and know-how that will promote the progress of science and useful arts, as the Constitution directs.³²

II. THE ABSTRACTION-FILTRATION-COMPARISON TEST

The Second Circuit's 1992 *Altai* decision was important for a number of reasons. For one thing, *Altai* recognized that the essentially utilitarian nature of computer programs meant that copyright law should be applied carefully to ensure that courts were not extending protection to functional aspects of programs, which should be free for all programmers to use (unless patented).³³ Second, *Altai* recognized that software developers are often constrained in their design decisions by, among other things, the need to be compatible with existing software or hardware, which should limit the scope of copyright protection in programs.³⁴ Third, the Second Circuit indicated in *Altai* that SSO was not a fruitful concept to employ when analyzing whether nonliteral elements of programs were within the scope of copyright protection.³⁵ Fourth, *Altai* rejected sweat-of-the-brow arguments for giving software a broad scope of protection because such arguments were inconsistent with Supreme Court precedent and fundamental principles of copyright law.³⁶ If software developers needed more legal protection than copyright could provide, the Second Circuit thought this was a matter for Congress.³⁷

Fifth and most important for the purposes of this section, *Altai* provided courts in subsequent cases with a more nuanced test for judging claims of software copyright infringement than *Whelan* and other prior cases had provided. The first step of *Altai*'s so-called "abstraction-filtration-comparison" test was to create a hierarchy of abstractions for the program alleged to be infringed, the second was to filter out unprotectable elements, and the third was to compare the remaining expression in the plaintiff's program with the defendant's program to determine whether the defendant infringed copyright.³⁸

In each respect, the *Altai* decision countered approaches that courts had taken in earlier cases. Section A discusses the cases to which *Altai* was, in

32. See U.S. CONST. art. I, § 8, cl. 8.

33. See *Comput. Assocs. Int'l, Inc. v. Altai, Inc.*, 982 F.2d 693, 703–05 (2d Cir. 1992).

34. *Id.* at 707–10.

35. *Id.* at 706.

36. *Id.* at 711–12 (citing *Feist Publ'ns, Inc. v. Rural Tel. Serv. Co.*, 499 U.S. 340, 349–50 (1991)).

37. *Id.* at 712.

38. See *id.* at 706–11.

a sense, responding. Section B provides some detail about the AFC test and why the Second Circuit thought many elements of programs should be filtered out before making final judgments about infringement. Section C explains what *Altai* got right and which aspects of the AFC test are in need of some refinement.

A. THE ROCKY ROAD TO *ALTAI*

In the early 1980s, U.S. appellate courts reviewed two software copyright cases in which the defendants had made exact copies of computer program object code.³⁹ Both defendants argued that the exact copying of the Apple II operating system (OS) programs was necessary to enable their computers to be compatible with the Apple II so that programs written to run on the Apple II could run on the defendants' platforms as well.⁴⁰ Apple prevailed in both cases.

Of the two cases, the Third Circuit's decision in *Apple Computer, Inc. v. Franklin Computer Corp.* merits discussion because of its unnuanced responses to the defendant's merger and § 102(b) arguments.⁴¹ Franklin contended that the ideas of the Apple II OS had merged with their expression because the only way Franklin could make its computers functionally compatible with the Apple II was by installing exact copies of the Apple OS programs on its machines.⁴² The Third Circuit rejected this argument, stating that achieving compatibility was "a commercial and competitive objective which does not enter into the somewhat metaphysical issue of whether particular ideas and expressions have merged."⁴³ Franklin also contended that the Apple OS programs were unprotectable by copyright law because § 102(b) excluded functional processes from the scope of copyright protection.⁴⁴ Section 102(b) provides that "[i]n no case does copyright protection . . . extend to any idea, procedure, process, system, method of operation, concept, principle or discovery, regardless of the form in which it is . . . embodied in . . . [the] work."⁴⁵ But the court

39. *Apple Comput., Inc. v. Formula Int'l, Inc.*, 725 F.2d 521 (9th Cir. 1984); *Apple Comput., Inc. v. Franklin Comput. Corp.*, 714 F.2d 1240 (3d Cir. 1983).

40. Formula's compatibility claim was more indirect than Franklin's, but Formula, like Franklin, was selling clones of the Apple II computer, and Formula objected to the issuance of an injunction because it would inhibit competitive entry into the computer market. *Formula*, 725 F.2d at 522–26; *Franklin*, 714 F.2d at 1253.

41. *Franklin*, 714 F.2d 1240.

42. *Id.* at 1253.

43. *Id.*

44. *Id.* at 1250–52.

45. 17 U.S.C. § 102(b).

rejected Franklin's argument because Congress had decided to treat programs as literary works.⁴⁶

The court in *Franklin* unquestionably reached the right result. If Congress' decision to extend copyright protection to computer programs was to be respected, it had to mean that exact copying of object code would get defendants in trouble, particularly where, as in *Franklin*, the copyist had not even tried to reimplement the Apple II OS functionality in different code.⁴⁷

The most troubling aspect of the *Franklin* decision was its strident rejection of compatibility as a possible justification for some copying from an existing program. Although eventually repudiated in several subsequent cases,⁴⁸ this dictum was recently revived in the CAFC's *Oracle* decision.⁴⁹ *Franklin* also took an unduly narrow view of § 102(b). The Third Circuit was right that § 102(b) should generally not be interpreted to allow the exact copying of object code just because such code is a "process."⁵⁰ But the court failed to acknowledge that § 102(b) excluded more than abstract ideas from the scope of copyright protection.⁵¹ The statutory exclusion of methods and processes embodied in programs had to be respected as well. Because of this exclusion, the scope of copyright protection in computer programs should be thinner than the scope of protection available to conventionally expressive works, such as novels and plays.⁵²

Although literal copying of program code, as in *Franklin*, presented an easy question for courts to answer, a more difficult question has been whether nonliteral elements of programs, such as SSO, qualify for copyright protection. That question was first addressed at the appellate level by the Third Circuit in *Whelan Associates, Inc. v. Jaslow Dental Lab., Inc.*⁵³

46. See *Franklin*, 714 F.2d at 1248–49, 1252–53.

47. *Id.* at 1245.

48. See *infra* text accompanying notes 259–264, 363.

49. *Oracle Am., Inc. v. Google Inc.*, 750 F.3d 1339, 1371 (Fed. Cir. 2014) (quoting the *Franklin* dicta).

50. *Franklin*, 714 F.2d at 1250–52. In rare cases, such as *Lexmark*, function and expression in a program may merge. *Lexmark Int'l, Inc. v. Static Control Components*, 387 F.3d 522, 540–42 (6th Cir. 2004). See *infra* text accompanying notes 318–322 for a discussion of function/expression merger.

51. See 17 U.S.C. § 102(b) (“In no case does copyright protection . . . extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery . . .”).

52. Congress added § 102(b) to the 1976 Act in part to ensure that courts would not give too broad an interpretation to software. See H.R. REP. NO. 94-1476 at 56–57 (1976), as reprinted in 1976 U.S.C.C.A.N. 5659, 5670.

53. 797 F.2d 1222 (3d Cir. 1986). A few District Court decisions prior to *Whelan* treated structural elements of programs as protectable expression. See, e.g., *SAS Inst., Inc.*

Jaslow had commissioned Whelan to develop a program to automate common business processes of dental laboratories.⁵⁴ The parties initially intended to exploit the software as partners,⁵⁵ but after a falling out, Jaslow decided to develop a program to do the same functions that would run on IBM-PCs.⁵⁶ Although Jaslow's and Whelan's programs were written in different programming languages and used different algorithms and data structures, the appellate court was impressed by similarities in the overall structures of the two programs,⁵⁷ in their file structures,⁵⁸ and in the way five modules performed functions.⁵⁹ Based on these similarities, the appellate court upheld a ruling that Jaslow infringed Whelan's copyright.⁶⁰

Jaslow's main defense was his contention that Congress had intended for copyright law to protect programs only against exact code copying.⁶¹

v. S&H Comput. Sys., 605 F. Supp. 816 (M.D. Tenn. 1985). *Whelan* relied in part on the *SAS* case. *See* 797 F.2d at 1239.

54. *Whelan Assoc., Inc. v. Jaslow Dental Lab., Inc.*, 609 F. Supp. 1307, 1309–10 (E.D. Pa. 1985). Jaslow had tried to develop this program on his own, but he lacked the skills to complete it, which is why he hired Whelan. *Id.*

55. *Id.* at 1312. Jaslow terminated this agreement and started his own firm to sell a competing product. *Id.* at 1313–15.

56. Whelan's contract explicitly provided that she would own IP rights in the Dentalab software. *Id.* at 1310. Despite this, Jaslow claimed to be the sole owner of the program (or at least a co-author of it) and licensed that software to third parties. *Id.* at 1315–17. The court held him as an infringer for the revenues he collected for selling Whelan's program to others. *Id.* at 1323.

57. 797 F.2d at 1239, 1247–48. In addition to selling infringing copies of Whelan's program, Jaslow, with the help of a contract programmer, had developed a competing program, which Whelan claimed infringed her copyright. 609 F. Supp. at 1314–15. This was the program that Whelan alleged had non-literally infringed her copyright. He called the new program "Dentlab," which the court held was confusingly similar to Whelan's "Dentalab" trademark. *Id.* at 1324.

58. *Whelan*, 797 F.2d at 1242–43.

59. *Id.* at 1245–46. The District Court had a difficult time understanding the technology at issue. Judge Vanartsdalen decided that Whelan's expert was more credible than Jaslow's expert because the former had observed the program in operation whereas the latter had only studied the source code for the two programs. *Whelan*, 609 F. Supp. at 1316. Vanartsdalen was also particularly impressed by the similarities in screen displays, *id.* at 1322, even though Whelan's claim of infringement was based on copying of structure from the underlying text of the program, not on screen displays. The District Court noted that Jaslow's program was not a translation from one computer language to another. *Id.* at 1315. The appellate court, on the other hand, was more precise about the structural similarities and recognized the risk of undue prejudice if courts used similarities in screen displays as evidence of similarities in program text, given that independently written code can produce the same results. 797 F.2d at 1244–45.

60. *Id.* at 1240–42.

61. *Id.* at 1235, 1241–42. *See also* Dennis S. Karjala, *Copyright Protection of Computer Documents, Reverse Engineering, and Professor Miller*, 19 U. DAYTON L. REV. 975, 984–89 (1994) (arguing that copyright should protect only against slavish copying of

The Third Circuit rejected this argument, reasoning that because copyright law had long protected structural elements of conventional literary works, such as novels and plays,⁶² copyright should protect the SSO of programs as well.⁶³ SSO included, in its view, “the manner in which the program operates, controls and regulates the computer in receiving, assembling, calculating, retaining, correlating, and producing useful information.”⁶⁴ Just as with novels and plays, anyone was free to copy the ideas from existing programs, but program SSO, *Whelan* announced, was protectable expression as long as there was more than one way to structure a program to achieve the program’s functions.⁶⁵ Because Jaslow could have used different SSO, his use of the same or similar ones as *Whelan*’s constituted infringement.⁶⁶

The Third Circuit acknowledged that software was a utilitarian work,⁶⁷ but it made no effort to distinguish between nonliteral elements of programs that should be regarded as protectable structures and those that might be unprotectable utilitarian processes.⁶⁸ *Whelan*’s test for infringement effectively rendered all program structure as protectable subject matter (unless there was truly no other way to structure the program). The Third Circuit noted that program structure and logic were “among the more significant costs in computer programming.”⁶⁹ The court concluded that without broad copyright protection for computer programs, there would be too little legal protection to provide proper incentives to invest in developing computer programs.⁷⁰

code). Jaslow infringed when he sold *Whelan*’s program, *see supra* note 56, but in light of later case developments, he may not have infringed as to the later-developed program. *See infra* note 450.

62. *See, e.g.*, *Sheldon v. Metro-Goldwyn Pictures Corp.*, 81 F.2d 49 (2d Cir. 1936) (finding infringement because of structural similarities between scenes in plaintiff’s play and defendant’s movie).

63. *Whelan*, 797 F.2d at 1233–34.

64. *Id.* at 1239–40 (quoting *Whelan*, 609 F. Supp. at 1320). This would seem to extend copyright protection to all program behavior, which is generally highly functional in character.

65. *Id.* at 1236; *see also id.* at 1224 n.1 (indicating that the opinion uses “‘structure,’ ‘sequence,’ and ‘organization’ interchangeably when referring to computer programs”).

66. *Whelan*, 797 F.2d at 1242–48.

67. *Id.* at 1236.

68. CONTU did not provide guidance on this score either. *See supra* text accompanying note 8.

69. *Whelan*, 797 F.2d at 1237.

70. *Id.* The Second Circuit in *Altai* noted that *Whelan* had been decided prior to the Supreme Court’s revocation of the sweat-of-the-brow doctrine in *Feist Publ’ns, Inc. v.*

Whelan and its test for infringement were initially followed in some cases,⁷¹ despite sharp criticisms in the law review literature for taking an overly expansive view of the scope of protection for programs.⁷² The *Whelan* test posited that only the general purpose or function of a program was an unprotectable idea under § 102(b), and all program SSO, no matter how abstract or standard it might be in the programming field, was protectable expression unless there was truly no alternative way to accomplish the function.⁷³ The first appellate court to break with *Whelan* was the Fifth Circuit Court of Appeals in *Plains Cotton Cooperative Ass'n v. Goodpasture Computer Service, Inc.*⁷⁴ The Fifth Circuit refused to follow *Whelan* because many of the structural similarities between user interfaces of the cotton trading programs at issue were “dictated by the externalities of the cotton market” and to be expected of a marketable program in that field.⁷⁵

Although *Whelan* did not involve program SSO necessary for achieving interoperability with other programs, coupled with *Franklin*'s rejection of compatibility defenses, the court's broad endorsement of SSO as protectable expression put unlicensed developers who reimplemented the interfaces of an existing program to make their own programs interoperable at risk of infringement.⁷⁶ Relying on the *Whelan* decision's endorsement of copyright protection for program SSO, Computer Associates (CA) brought a lawsuit against Altai for nonliteral copyright infringement that came before the Second Circuit in 1992.⁷⁷ Altai defended by arguing that it was

Rural Tel. Serv. Co., 499 U.S. 340, 349–50 (1991). *Comput. Assocs. Int'l, Inc. v. Altai, Inc.*, 982 F.2d 693, 711–12 (2d Cir. 1992).

71. See, e.g., *Broderbund Software, Inc. v. Unison World, Inc.*, 648 F. Supp. 1127, 1133 (N.D. Cal. 1986).

72. See *Altai*, 982 F.2d at 711–12 (citing five critics of the *Whelan* decision); Englund, *supra* note 10, at 875–76 (noting that not just program ideas, but also program processes, should be excluded from the scope of copyright protection). See also *infra* notes 82–85 and accompanying text for a fuller account of *Altai*'s criticism of *Whelan*.

73. *Whelan*, 797 F.2d at 1236.

74. 807 F.2d 1256 (5th Cir. 1987).

75. *Id.* at 1262.

76. One post-*Whelan* District Court upheld a second-comer's reuse of a first-comer's variation on an interface protocol that was necessary to achieve compatibility. See *Secure Servs., Inc. v. Time & Space Proc'g*, 722 F. Supp. 1354 (E.D. Va. 1989) (finding no infringement to re-implement digital handshake for secure fax machines to be sold to the government because variations lacked sufficient originality). *Whelan* was not cited.

77. See Reply Brief for Plaintiff-Appellant at 8–18, *Comput. Assocs. Int'l, Inc. v. Altai, Inc.*, 982 F.2d 693 (2d Cir. 1992) (No. 91-7893), 1991 WL 11010234 (relying heavily on *Whelan*).

necessary to use the same SSO in its program in order to be compatible with third-party programs.⁷⁸

B. THE *Altai* DECISION AND THE AFC TEST

Computer Associates Int'l, Inc. v. Altai, Inc. was the first appellate court decision to consider a compatibility defense to a claim of nonliteral copyright infringement.⁷⁹ CA charged Altai with copyright infringement for copying program SSO—overall structure, macros, lists of services, and parameter lists (i.e., lists of information that needed to be sent and received by subroutines of the affected programs)—from CA's scheduling program.⁸⁰ Altai argued that it needed to use the same parameter lists and macros because this SSO was necessary for its programs to be compatible with the IBM computers on which both its own and CA's scheduling programs were designed to run; other similarities were, moreover, to be expected in programs of that kind.⁸¹

Before setting forth its alternative analysis, the Second Circuit discussed the *Whelan* decision at some length. The *Altai* court agreed with the Third Circuit that nonliteral elements of programs could be protected by copyright,⁸² but it did not find *Whelan*'s SSO concept helpful in distinguishing between which nonliteral elements of programs were protectable by copyright law and which were not. The court regarded the SSO concept as “demonstrat[ing] a flawed understanding of a computer program's method of operation” and as resting on a “somewhat outdated appreciation of computer science.”⁸³ While it characterized *Whelan* as “the most thoughtful” attempt to apply the idea/expression distinction to computer programs, the court also noted the widespread criticism of the decision as “conceptually overbroad.”⁸⁴ The *Whelan* test for infringement of program SSO was, moreover, “descriptively inadequate,” relying “too

78. Brief of Defendant-Appellee at 10–13, *Comput. Assoc. Int'l, Inc. v. Altai, Inc.*, 982 F.2d 693 (2d Cir. 1992) (No. 91-7893), 1991 WL 11010233.

79. *Comput. Assocs. Int'l, Inc. v. Altai, Inc.*, 775 F. Supp. 544 (E.D.N.Y. 1991), *aff'd*, 982 F.2d 693 (2d Cir. 1992).

80. CA sued Altai for copyright infringement after learning that Altai's program contained code directly copied from its CA-Scheduler program by one of CA's former employees whom Altai had hired. Altai purged the tainted code from its program and assigned a clean-room team of programmers to reimplement the compatibility components in new non-infringing code. But CA argued that there were still substantial similarities in SSO that Altai had copied from its program. *Altai*, 982 F.2d at 697–700.

81. *See id.* at 714–15.

82. *Id.* at 702.

83. *Id.* at 706.

84. *Id.* at 705.

heavily on metaphysical distinctions” instead of on “practical considerations.”⁸⁵

The “essentially utilitarian nature of a computer program,” the court said in *Altai*, makes distinguishing program ideas and expressions more difficult because compared to aesthetic works, “computer programs hover even more closely to the elusive boundary line described in § 102(b).”⁸⁶ The Second Circuit characterized the Supreme Court’s decision in *Baker v. Selden* as the “doctrinal starting point” for analyzing the scope of copyright in “utilitarian works” because that decision recognized that the copyright in a work describing or illustrating a useful art did not extend to that art.⁸⁷ The Second Circuit quoted at length from *Baker* and its holding that Selden’s bookkeeping system and the forms that illustrated its use were not protectable by copyright law.⁸⁸ The court perceived computer programs to be “roughly analogous” to Selden’s book, and consistent with *Baker*, it stated “those elements of a computer program that are necessarily incidental to its function are similarly unprotectable.”⁸⁹

To ensure that unprotectable elements of programs would not be inadvertently swept into the infringement determination, the Second Circuit thought it important to fashion a new test for software copyright infringement that would filter those elements out. It announced a three-step test to achieve this objective in cases involving nonliteral software copyright infringement.⁹⁰ The first step required constructing a hierarchy of abstractions for the program alleged to be infringed. The second step required filtering out the unprotectable elements of programs. The court said filtration should exclude: aspects of programs that are dictated by efficiency, design choices that are constrained by external factors, and elements of programs that are in the public domain, such as commonplace programming techniques, ideas, and know-how.⁹¹ The *Altai* test’s third step called for comparing the “golden nugget[s]” of expression remaining after filtration to determine if there was substantial similarity in the nonliteral expression that the defendant had copied from the plaintiff’s program.⁹²

Applying this test, the court in *Altai* was satisfied that some similarities between CA’s and *Altai*’s programs were due to public domain elements,

85. *Id.* at 705–06.

86. *Id.* at 704.

87. *Id.*

88. *Id.* at 704–05 (quoting *Baker v. Selden*, 101 U.S. 99, 103 (1879)).

89. *Id.*

90. *See id.* at 706–11.

91. *Id.* at 707–10.

92. *Id.* at 710.

some were at too high a level of generality, and others were “dictated by the functional demands of the program.”⁹³ The parameter list similarities, in particular, were seen as necessary to develop a program that would be compatible with the IBM systems.⁹⁴

The Second Circuit recognized that applying the AFC test to software might mean that copyright would “serve[] as a relatively weak barrier against public access to the theoretical interstices” of program design, but that “results from the hybrid nature of a computer program, which, while it is literary expression, is also a highly functional, utilitarian component in the larger process of computing.”⁹⁵ The AFC test “not only comports with, but advances the constitutional policies underlying the Copyright Act,” the court wrote.⁹⁶ CA’s economic arguments in favor of broad copyright protection for program SSO were deemed inconsistent with Supreme Court precedents.⁹⁷ Adopting CA’s theory would, the court added, “have a corrosive effect on certain fundamental tenets of copyright doctrine.”⁹⁸ Copyright law, the Second Circuit recognized, should not be construed to give utility patent-like protection to highly functional program SSO.⁹⁹

C. A CLOSER LOOK AT *ALTAI*’S FILTRATION FACTORS

The *Altai* decision identified three categories of nonliteral elements of computer programs that should be filtered out of consideration during infringement analysis. The first was “elements dictated by efficiency.”¹⁰⁰ Given that *Altai* did not raise an efficiency defense, the Second Circuit did not need to discuss the exclusion of efficient design elements from the scope of copyright. However, the court was influenced by the “successive filtering” test for software copyright infringement proposed in the Nimmer treatise, which identified efficient program designs as unprotectable elements.¹⁰¹ The Second Circuit relied upon the merger doctrine to justify

93. *Id.* at 714–15 (quoting Judge Pratt’s District Court opinion).

94. *Id.*

95. *Id.* at 712.

96. *Id.* at 711.

97. *Id.* at 711–12 (citing *Feist Publ’ns., Inc. v. Rural Tel. Serv. Co.*, 499 U.S. 340, 349–50 (1991)) (rejecting “sweat of the brow” copyright claim in white pages listings of a telephone directory).

98. *Id.* at 712.

99. *Id.*; see also *infra* Part V (discussing why copyrights should not be interpreted to give patent-like protection to program innovations).

100. *Altai*, 982 F.2d at 707–09. Efficiency as a constraint is discussed further *infra* text accompanying notes 344–347.

101. See *Altai*, 982 F.2d. at 707 (citing *NIMMER ON COPYRIGHT*, *supra* note 10, at 13.03 [F]). The heart of the Nimmer successive filtering test became known as the AFC test after

the filtration of efficient designs insofar as those nonliteral elements of a program were “dictated by considerations of efficiency, so as to be necessarily incidental to that idea.”¹⁰² By invoking the merger doctrine to exclude efficient elements of programs, the Second Circuit remained faithful to *Baker v. Selden* and the Supreme Court’s holding that functional design elements in copyrighted works are not within the scope of copyright protection, although they may be eligible for patenting.

Altai’s exclusion of efficient design elements was an important pronouncement because cases such as *Whelan* had not considered efficiency as a limiting principle in software cases. Under the Third Circuit’s conception of software copyright protection in *Whelan*, the question was whether there was *any* other way to carry out a function, and if there was, then the second comer had to do it another way and not copy the plaintiff’s SSO.¹⁰³ The Second Circuit, however, noted that “[i]n the context of computer program design, the concept of efficiency is akin to deriving the most concise logical proof or formulating the most succinct mathematical computation.”¹⁰⁴ There might well be, *ex ante*, myriad ways to carry out some functions in a program, but “efficiency concerns may so narrow the practical range of choice as to make only one or two forms of expression workable options.”¹⁰⁵ The court was reluctant to force programmers to use inefficient (and hence inferior) program designs when efficient ones were available.¹⁰⁶

the *Altai* decision. The court also cited to Professor Menell, *supra* note 10, at 1052, when discussing efficiency as a limiting principle on copyright.

102. *Altai*, 982 F.2d at 707. The “necessary incidents” language, of course, traces back to *Baker*. See 101 U.S. 99, 103 (1879).

103. *Whelan Associates Inc. v. Jaslow Dental Lab., Inc.*, 797 F.2d 1222, 1236 (3d Cir. 1986). The Third Circuit noted that efficiency is a “prime concern” of programmers and the more efficient a program was, the more valuable it would be. *Id.* at 1230. But the court did not perceive efficiency as a design constraint. See also *Lotus Dev. Corp. v. Paperback Int’l, Inc.*, 740 F. Supp. 37, 57 (D. Mass. 1990) (rejecting arguments that the functionality of an interface design should limit the scope of copyright protection).

104. *Altai*, 982 F.2d at 708. Computer science books teach programmers about the relative efficiencies of different algorithms and other program design elements for particular types of computations. See generally DONALD KNUTH, *THE ART OF COMPUTER PROGRAMMING: FUNDAMENTAL ALGORITHMS* (3d ed. 1997).

105. *Altai*, 982 F.2d at 708.

106. See *id.* The efficiency exclusion, endorsed in *Altai*, has been influential in subsequent cases. See, e.g., *Bay State Tech., Inc. v. Bentley Sys., Inc.*, 946 F. Supp. 1079, 1088 (D. Mass. 1996); see also *Zalewski v. Cicero Builder Dev. Inc.*, 754 F.3d 95, 105 (2d Cir. 2014) (efficiency considerations limit scope of copyright in architectural works).

The second filtration category concerned those “elements dictated by external factors.”¹⁰⁷ Relying heavily on the Nimmer treatise, *Altai* identified five types of “extrinsic considerations” that may “circumscribe[]” a programmer’s “freedom of design choice.”¹⁰⁸ They were:

- (1) the mechanical specifications of the computer on which a particular program is intended to run;
- (2) compatibility requirements of other programs with which a program is intended to run;
- (3) computer manufacturers’ design standards;
- (4) demands of the industry being serviced; and
- (5) widely accepted programming practices within the computer industry.¹⁰⁹

Although “dictated by” is the language of merger, as *Altai* recognized in its discussion of efficiency, the Second Circuit pointed to the scenes a faire doctrine as its main doctrinal justification for the “external factors” category.¹¹⁰ This doctrine excludes from the scope of copyright protection elements that are inevitable or commonly present in works of that kind.¹¹¹ Scenes a faire had, the court noted, been employed to limit copyright liability in some prior computer software cases.¹¹² When applying the AFC test to the facts in *Altai*, the Second Circuit viewed similarities in the organizational charts of the two programs as scenes a faire elements because they “follow[ed] naturally from the work’s theme rather than from the author’s creativity.”¹¹³

It does seem appropriate to characterize widely accepted programming practices or similarities driven by the demands of the industry as falling within the scenes a faire doctrine.¹¹⁴ The merger doctrine, however, is more

107. *Altai*, 982 F.2d at 709–10.

108. *Id.* at 709.

109. *Id.* at 709–10 (citing NIMMER ON COPYRIGHT, *supra* note 10, at 13-66-71).

110. *Altai*, 982 F.2d at 709–10. The scenes a faire doctrine is distinct from the merger doctrine because the latter focuses constraints on the range of expressive choices available to authors. See *infra* Part IV for an extended discussion of the merger doctrine.

111. *Altai*, 982 F.2d at 709 (citing *Hoehling v. Universal City Studios, Inc.*, 618 F.2d 972, 979 (2d Cir. 1980)) (similarities between book and movie about the Nazi era in Germany were unprotectable scenes a faire elements).

112. *Id.* at 709–10 (citing to four prior cases, including *Data East USA, Inc. v. Epyx, Inc.*, 862 F.2d 204 (9th Cir. 1998)) (similarities in karate videogames were unprotectable as to scenes a faire elements).

113. *Id.* at 715.

114. See *id.* at 710. The Second Circuit pointed to *Plains Cotton Coop. Ass’n v. Goodpasture Comput. Serv., Inc.*, 807 F.2d 1256, 1262 (5th Cir. 1987), which had ruled that similarities in programs were due to demands of the cotton market, to illustrate the point. The Fifth Circuit in *Plains Cotton*, however, did not mention the scenes a faire doctrine in its decision; instead it said that similarities were “dictated by” demands of the industry. *Id.* “Dictated by” is the language of merger.

pertinent to constraints imposed by mechanical specifications of the computer on which a program is run, computer manufacturer design standards, and requirements for achieving compatibility with other programs because these are “necessary incidents” constraints.¹¹⁵ In applying the AFC test, the Second Circuit quoted the District Court as having found that some similarities between CA’s and *Altai*’s programs were “‘dictated by the functional demands of the program.’”¹¹⁶ Dictated by, as noted above, is the language of merger, not of scenes a faire.

The third category of unprotectables consisted of “elements taken from the public domain.”¹¹⁷ This might have seemed unnecessary to say, but perhaps the Second Circuit wanted to make sure that courts would be on the lookout for public domain elements. The *Whelan* decision had been so sweeping in its conception of the scope of copyright in software—every bit of program SSO was said to be protectable expression unless there were no alternative choices possible¹¹⁸—that the *Altai* court’s reminder that programs, like other works, contain public domain elements was useful. The public domain category of filtration has been used in some post-*Altai* cases,¹¹⁹ but the Second Circuit did not spell out what kinds of public domain elements it thought should be filtered out.¹²⁰

What is strangely missing from *Altai*’s list of unprotectable elements in computer programs that must be filtered out before deciding whether

115. Some subsequent cases have treated compatibility components of programs as unprotectable under the merger doctrine. *See, e.g.,* Lexmark Int’l, Inc. v. Static Control Components, 387 F.3d 522, 540–42 (6th Cir. 2004); *see also* Bateman v. Mnemonics Inc., 79 F.3d 1532, 1544–48 (11th Cir. 1996) (reversing lower court for failure to give proper jury instruction about the possible need to use the same interface code to attain compatibility).

116. *Altai*, 982 F.2d at 714 (quoting *Comput. Assocs. Int’l, Inc. v. Altai, Inc.*, 775 F. Supp. 544, 562 (E.D.N.Y. 1991)).

117. *Id.* at 710.

118. *Whelan Associates Inc. v. Jaslow Dental Lab., Inc.*, 797 F.2d 1222, 1236 (3d Cir. 1986).

119. *See, e.g.,* Paycom Payroll LLC v. Richison, 758 F.3d 1198, 1205 (10th Cir. 2014) (directing filtration of “ideas, processes, facts, public domain information, merger material, scenes a faire material, and other unprotected elements suggested by the particular facts of the program under examination”); *Gates Rubber Co. v. Bando Chem. Indus.*, 9 F.3d 823, 842–43 (10th Cir. 1993) (defendant could lawfully reuse the plaintiff’s mathematical constants because they were in the public domain as facts).

120. In support of the public domain category, *Altai* cited *Brown Bag Software v. Symantec Software*, 960 F.2d 1465, 1473 (9th Cir. 1992) (“Plaintiffs may not claim copyright protection of an expression that is, if not standard, then commonplace in the computer software industry.”). This quote would have more suitably illustrated a type 5 external factors excludable under the scenes a faire doctrine in the *Altai* conception of the filtration categories.

infringement has occurred are the categories that § 102(b) renders unprotectable by copyright law: procedures, processes, systems, and methods of operation. The Second Circuit quoted that provision once and cited it in three places.¹²¹ But curiously, it did not consider what kinds of procedures or methods should be unprotectable aspects of programs, let alone direct the filtration of these § 102(b) elements. Later cases have identified algorithms and functional behavior as among the structural elements of programs that must be excluded from protection under § 102(b).¹²²

The Second Circuit may have thought the exclusion of procedures, etc., was unnecessary because it regarded the unprotectable elements it did identify as proxies for the procedure, process, system, and method of operation exclusions that § 102(b) says are unprotectable. But merger, scenes a faire, and public domain are different types of limiting principles than processes and methods of operation. If the Second Circuit regarded efficiency, external factors, and public domain elements as proxies for the § 102(b) excludables, it should have explained why the words of the statute should be ignored and why these proxies were appropriate.

It is also possible that the Second Circuit did not perceive a need to filter out these § 102(b) excludables because the *Altai* case arguably did not involve any claim about processes or methods of operation. However, other cases have identified the functional requirements for achieving interoperability with other programs as unprotectable procedures under § 102(b).¹²³ So this does not explain the omission.

A third possibility—and probably the true explanation—is that the Second Circuit was relying on the Nimmer treatise's successive filtering method for judging software copyright infringement. Because that treatise did not identify procedures, processes, systems, or methods of operation as excludable elements in computer programs, neither did the Second Circuit. The Nimmer treatise has systematically deflected attention away from the wider meaning of § 102(b), treating it as merely a restatement of the idea/expression distinction.¹²⁴ Under the influence of the Nimmer treatise,

121. *Altai*, 782 F.2d at 703–04.

122. See *infra* notes 135–42 and accompanying text.

123. See, e.g., *Sega Enters. Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1522, 1526 (9th Cir. 1992).

124. See Pamela Samuelson, *Why Copyright Excludes Systems and Processes From the Scope of Its Protection*, 85 TEX. L. REV. 1921, 1953–61 (2007). The Nimmer treatise has ignored the myriad cases that, under the influence of *Baker v. Selden*, have held such things as procedures, processes, systems, and methods of operation are unprotectable elements of copyrighted works. *Id.* at 1936–44 (discussing the cases). Congress intended

courts have sometimes been reluctant to pay attention and give content to these exclusions, even though these elements have a wider range of possible applications for computer programs than for any other type of copyrighted work.¹²⁵ Procedures, processes, systems, and methods of operation are, after all, functional design elements of the “essentially utilitarian nature of computer program[s].”¹²⁶

III. CONCEPTUALIZING THE PROPER ROLE OF § 102(b) IN COMPUTER PROGRAM COPYRIGHT CASES

Because computer programs embody many procedures, processes, systems and methods of operation that lie beyond the scope of copyright, it would be logical for § 102(b) to have a significant, and perhaps even a central, role in the analysis of claims of software copyright infringement. One way to accomplish this objective would be to adapt the *Altai* AFC test, as some courts have already done, by adding a fourth category of program design elements derived from § 102(b) that should be filtered out before proceeding with the final stage of infringement analysis.¹²⁷ Another way would be, in appropriate cases, to consider and apply the functional design element exclusions in § 102(b) without tying them to the *Altai* framework, which other courts have also already done.¹²⁸

Section A discusses five propositions about § 102(b) that should be uncontroversial and shows that courts in recent years have been more receptive to employing § 102(b) as a limiting principle of software copyright law. Sections B and C review the *Lotus v. Borland* and *Oracle v. Google* cases, and explain why the application of § 102(b) in those cases has generated some controversy. It is unfortunate that the Supreme Court split 4-4 in *Borland* in 1995 and that twenty years later, it declined to hear Google’s appeal; the split in the circuits as to the proper interpretation of § 102(b) is unlikely to be resolved any time soon. Section D explains why

to codify these common law exclusions from the scope of copyright by including § 102(b) in the statute *Id.* at 1944–52 (discussing the legislative history of § 102(b)); *see also* Ann Bartow, *The Hegemony of the Copyright Treatise*, 73 U. CINN. L. REV. 581 (2004) (criticizing courts for relying heavily on copyright treatises, including Nimmer’s, in interpreting copyright law).

125. The *Franklin* and *Whelan* decisions are exemplary of this trend. *See* Apple Comput., Inc. v. Franklin Comput. Corp., 714 F.2d 1240, 1250–52 (3d Cir. 1983); *Whelan Associates Inc. v. Jaslow Dental Lab., Inc.*, 797 F.2d 1222, 1234–36 (3d Cir. 1986).

126. *Altai*, 982 F.2d at 704.

127. *See, e.g.*, *Gates Rubber Co. v. Bando Chem. Indus.*, 9 F.3d 823, 836–37 (10th Cir. 1993); *see also infra* text accompanying note 453.

128. *See, e.g.*, *Lotus Dev. Corp. v. Borland Int’l, Inc.*, 49 F.3d 807, 815–16 (1st Cir. 1995), *aff’d by an equally divided Court*, 516 U.S. 233 (1996).

the Federal Circuit's interpretation of § 102(b) compatibility defenses in *Oracle* should be repudiated in subsequent software copyright cases.

A. FIVE UNCONTROVERSIAL PROPOSITIONS ABOUT § 102(b)

This Section articulates five propositions about § 102(b) that should be uncontroversial. First, § 102(b) should not be interpreted in a manner that would deprive programs of any copyright protection. Second, the procedure, process, system, and method of operation exclusions in § 102(b) are meaningful limits on the scope of copyright protection available to programs. Third, one function of the § 102(b) exclusion of processes and methods from the scope of copyright is to maintain boundaries between the copyright and patent regimes. Fourth, the utilitarian nature of programs differentiates them from conventional literary works because they contain functional design elements such as processes that are excluded under § 102(b). Fifth, "SSO" is not a useful way to distinguish between those nonliteral elements of programs that are unprotectable under § 102(b) and those that constitute protectable expressions.

1. *Section 102(b) Does Not Exclude Program Code from Protection*

The least controversial proposition about § 102(b) in relation to computer programs is that Congress could not possibly have intended courts to give a completely literal interpretation to § 102(b) because this would render programs ineligible for copyright protection. Object code is unquestionably a functional process. Therefore, the Third Circuit in *Franklin* correctly rejected Franklin's § 102(b) defense: Franklin copied the Apple OS programs, bit for bit, and did not even try to reimplement the functionality of the Apple programs in independently written code.¹²⁹ Furthermore, while object code is the most functional embodiment of program processes, because source code forms of programs function as detailed statements and instructions for carrying out certain tasks, they are unquestionably "procedures" within the normal meaning of that word. Yet, to respect Congress' decision to extend copyright protection to computer programs, original program code, whether in source or object code form, should generally be protectable by copyright law.¹³⁰

129. See *supra* note 47 and accompanying text.

130. Occasionally, a program may either have insufficient originality to support a copyright or be rendered unprotectable because function and expression have merged. See, e.g., *Lexmark Int'l, Inc. v. Static Control Components*, 387 F.3d 522, 540–42 (6th Cir. 2004) (questioning the originality of a program embedded in a printer cartridge, but also applying the merger doctrine to it).

2. *The Procedure, Process, System and Method of Operation Exclusions of § 102(b) Must Mean Something*

A second uncontroversial proposition is that the words of exclusion from the scope of copyright protection for “procedure[s], process[es], system[s], [and] method[s] of operation” in § 102(b) must mean something.¹³¹ As a matter of logic, they cannot be synonymous with the “idea” exclusion because “idea” is but one of eight categories of excludable elements listed in § 102(b). One term cannot subsume the other seven. Giving meaning to each of the statutory exclusions is consistent with conventional canons of statutory construction.¹³² When a statute specifically identifies several categories of unprotectable elements and says that “[i]n no case” should any of them be within the scope of copyright protection,¹³³ courts should not read all but one of the terms out of the statute, as the appellate courts did in *Whelan* and *Oracle*.

That these functional design elements excluded by § 102(b) should be given close attention in software copyright infringement cases is also evident from the legislative history of the 1976 Act. Both the House and Senate Reports explained why § 102(b) was put in the statute:

Some concern has been expressed lest copyright in computer programs should extend protection to the *methodology or processes* adopted by the programmer, rather than merely to the “writing” expressing his ideas. Section 102(b) is intended, among other things, to make clear that the expression adopted by the programmer is the copyrightable element in a computer program, and that the *actual processes or methods* embodied in the program are not within the scope of the copyright law.¹³⁴

Courts in software copyright cases have identified several types of nonliteral elements of programs as unprotectable procedures, processes, systems, or methods of operation under § 102(b). Among the elements that

131. 17 U.S.C. § 102(b).

132. See *Montclair v. Ramsdell*, 107 U.S. 147, 152 (1883) (“It is the duty of the court to give effect, if possible, to every clause and word of a statute, avoiding, if it may be, any construction which implies that the legislature was ignorant of the meaning of the language it employed.”); LARRY M. EIG, CONG. RESEARCH SERV., 97-589, STATUTORY INTERPRETATION: GENERAL PRINCIPLES AND RECENT TRENDS 13 (2011) (“The modern variant [of this principle] is that statutes should be construed ‘so as to avoid rendering superfluous’ any statutory language” (quoting *Hibbs v. Winn*, 542 U.S. 88, 101 (2004))).

133. 17 U.S.C. § 102(b).

134. H.R. REP. NO. 94-1476, at 57 (1976); S. REP. NO. 94-473, at 54 (1975) (emphasis added).

courts have filtered out are algorithms,¹³⁵ mathematical constants,¹³⁶ rules editing methods,¹³⁷ methods of calculation,¹³⁸ command structures,¹³⁹ data structures,¹⁴⁰ interfaces necessary to interoperability,¹⁴¹ and functional program behavior.¹⁴² Also noteworthy is the Copyright Office's articulation of functional elements of computer programs that the Office regards as unprotectable by copyright law.¹⁴³

The § 102(b) exclusions apply, of course, to all types of copyrighted works, not just to computer programs. Procedures, processes, systems and methods of operation have, in fact, been excluded from the scope of copyright protection both before and after § 102(b) was added to the statute. Consider, for example, *Brief English Systems v. Owen* which illustrates the system exclusion.¹⁴⁴ The Second Circuit decided that Owen was free to write his own book on the shorthand system that the plaintiff had devised because that system was ineligible for copyright protection.¹⁴⁵ *Taylor Instrument Cos. v. Fawley-Brost Co.* illustrates the method of operation

135. See, e.g., *Torah Soft Ltd. v. Drosnin*, 136 F. Supp. 2d 276, 291 (S.D.N.Y. 2001).

136. See, e.g., *Gates Rubber Co. v. Bando Chem. Indus.*, 9 F.3d 823, 842–43 (10th Cir. 1993). *Gates Rubber* directs use of a “process/expression” distinction in computer program cases. *Id.*

137. See, e.g., *Ilog v. Bell Logic, LLC*, 181 F. Supp. 2d 3, 14 (D. Mass. 2002).

138. See, e.g., *Harbor Software v. Applied Sys., Inc.*, 925 F. Supp. 1042, 1052 (S.D.N.Y. 1996).

139. See, e.g., *Mitek Holdings, Inc. v. Arce Eng'g Co.*, 89 F.3d 1548, 1557 (11th Cir. 1996).

140. See, e.g., *Baystate Techs., Inc. v. Bentley Systems, Inc.*, 946 F. Supp. 1079, 1088–89 (D. Mass. 1996).

141. See, e.g., *Sega Enters. Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1522 (9th Cir. 1992).

142. See, e.g., *O.P. Solutions v. Intell. Prop. Network Ltd.*, No. 96 Civ. 7952 (LAP), 1999 WL 47191, at *16–20 (S.D.N.Y. Feb. 2, 1999).

143. Compendium of U.S. Copyright Practices (3d ed. 2015) [hereinafter *Compendium*], § 721.7 (“[T]he Office will not register the functional aspects of a computer program, such as the program’s algorithm, formatting, functions, logics, system design, and the like.”); see also *id.* at § 721.9(J) (indicating that registration will not be accepted if the program author claims copyright in an algorithm, computation, data, formatting, formulas, interfaces, language, layout, logic, menu screens, models, organization, protocols, and system design, among other categories of unprotectable elements).

144. 48 F.2d 555 (2d Cir. 1931).

145. *Id.* at 556 (“There is no literary merit in a mere system of condensing written words into less than the number of letters usually used to spell them out. Copyrightable material is found, if at all, in the explanation of how to do it.”). Numbering systems for hardware parts are similarly unprotectable by copyright law. See, e.g., *ATC Distrib. Group, Inc. v. Whatever It Takes Transmission & Parts, Inc.*, 402 F.3d 700 (6th Cir. 2005); *Southco, Inc. v. Kanebridge Corp.*, 390 F.3d 276 (3d Cir. 2004).

exclusion.¹⁴⁶ The Seventh Circuit rejected Taylor’s claim of copyright in charts designed for use as part of a method of operating a temperature recording machine.¹⁴⁷ *Publications Int’l Ltd. v. Meredith Corp.* illustrates the exclusion of procedures from the scope of copyright.¹⁴⁸ The Seventh Circuit held that individual recipes were unprotectable procedures under § 102(b), and others could freely copy them.¹⁴⁹ *Bikram’s Yoga College of India, L.P. v. Evolation Yoga, LLC*, illustrates the process exclusion.¹⁵⁰ The Ninth Circuit held that the series of yoga poses and breathing exercises that Bikram Choudhury had developed were unprotectable processes under § 102(b) because of their functional character in helping to attaining psychological and spiritual well-being.¹⁵¹

3. *The Process and System Exclusions of § 102(b) Are Partly Aimed at Maintaining Boundaries between Copyright and Patent Laws*

A third proposition about § 102(b) that should be uncontroversial is that procedures, processes, systems, and methods of operation are excluded from the scope of copyright protection in part to maintain a balance between the role of copyright in protecting authorial expression and the role of patent law in protecting inventions in the useful arts.¹⁵² Patented processes, systems and methods of operation are often described in documents or

146. 139 F.2d 98 (7th Cir. 1943); *see also* *Brown Instrument Co. v. Warner*, 161 F.2d 910, 911 (D.C. Cir. 1947) (upholding the Copyright Office’s refusal to register charts for recording data).

147. *Taylor Instrument*, 139 F.2d at 100–01; *see also* *Coates-Freeman Assocs., Inc. v. Polaroid Corp.* 792 F. Supp. 879, 884–85 (D. Mass. 1992) (problem-solving method unprotectable by copyright law).

148. 88 F.3d 473 (7th Cir. 1996).

149. *Id.* at 480–81.

150. 803 F.3d 1032 (9th Cir. 2015); *see also* *Palmer v. Braun*, 287 F.3d 1325, 1333–34 (11th Cir. 2002) (citing to § 102(b) and affirming denial of a preliminary injunction because similarities between Palmer’s and Braun’s courses and course materials were largely due to the fact that they were teaching the same processes for raising human consciousness).

151. *Bikram’s Yoga*, 803 F.3d at 1039–40.

152. *See, e.g.,* *Incredible Techs., Inc. v. Virtual Techs., Inc.*, 284 F. Supp. 2d 1069, 1078 (N.D. Ill. 2003), *aff’d*, 400 F.3d 1007 (7th Cir. 2015) (control panel features of videogame were “potentially patentable-but not copyrightable”); *Bateman v. Mnemonics, Inc.*, 79 F.3d 1532, 1541, 1546 (11th Cir. 1996); *Gates Rubber Co. v. Bando Chem. Indus.*, 9 F.3d 823, 837 (10th Cir. 1993) (recognizing that some program processes may be patentable). Processes are one of four categories of statutory subject matters eligible for patent. *See* 35 U.S.C. § 101. Claims for patentable processes are often stated as methods for carrying out certain operations. Patents for machines are typically stated as systems to perform certain functions.

illustrated in drawings. But “the principle is the same in all,” as the Supreme Court said in *Baker v. Selden* more than a hundred twenty years ago.¹⁵³ “The description of the [useful] art in a book, though entitled to the benefit of copyright, lays no foundation for an exclusive claim to the art itself [That] can only be secured, if it can be secured at all, by letters-patent.”¹⁵⁴ It would be “a surprise and fraud on the public,” the Supreme Court wrote, if Selden could get through the copyright in his book a longer duration of exclusive rights than if he had been able to get the patent he sought (which he apparently failed to obtain).¹⁵⁵

In *Taylor Instrument*, the Seventh Circuit invoked *Baker* in deciding that Taylor’s charts for recording temperatures over time were not proper subject matter for copyright protection.¹⁵⁶ Although Taylor had obtained registration certificates from the Copyright Office, the charts were unprotectable by copyright law because they were indispensable parts of Taylor’s recording machines.¹⁵⁷ Fawley-Brost, the alleged infringer, was entitled to provide a competitive alternative to those customers who owned Taylor’s temperature recording machines and wanted cheaper charts that would interoperate with the Taylor machines. The Seventh Circuit observed:

While it may be difficult to determine in which field [of intellectual property] protection must be sought, it is plain, so we think, that it must be in one [copyright] or the other [patent]; it cannot be found in both. In other words, there is no overlapping territory, even though the line of separation may in some instances be difficult of exact ascertainment.¹⁵⁸

153. *Baker v. Selden*, 101 U.S. 99 (1879).

154. *Id.* at 105; see also *Bikram’s Yoga*, 803 F.3d at 1039–40 (ruling yoga healing methodology is uncopyrightable, but stating that if it is “entitled to protection at all, that protection is more properly sought through the patent process”).

155. *Id.*; see Pamela Samuelson, *Baker v. Selden: Sharpening the Distinction Between Authorship and Invention*, in *INTELLECTUAL PROPERTY STORIES* 160–61 (Rochelle C. Dreyfuss & Jane C. Ginsburg eds., 2005) (noting that the preface to Selden’s book referred to his patent application). Had the Court accepted Selden’s claim of copyright protection in the bookkeeping system, he could have had exclusive rights in it for up to forty-two years instead of the much shorter duration available had a patent issued for the system.

156. 139 F.2d 98, 99–100 (7th Cir. 1943).

157. *Id.* at 100.

158. *Id.* at 99; see also *Brown Instrument Co. v. Warner*, 161 F.2d 910, 911 (D.C. Cir. 1947) (only patent, not copyright, protection was available to charts as parts of recording machines). Note that *Taylor Instrument* and *Brown Instrument* are pre-software interoperability cases.

Similar considerations underlay the Ninth Circuit decision to reject Sega's infringement claim in *Sega Enters. Ltd. v. Accolade, Inc.*¹⁵⁹ The infringement suit commenced after Accolade reverse engineered Sega programs to get access to information about the functional requirements for achieving interoperability with Sega's Genesis platform. Sega sued to stop Accolade, arguing that the copies made in the course of reverse engineering were infringements.¹⁶⁰ The Ninth Circuit stated that "[i]f disassembly of copyrighted object code is per se an unfair use, the owner of the copyright gains a de facto monopoly over the functional aspects of his work—aspects that were expressly denied copyright protection by Congress" under § 102(b).¹⁶¹ The court characterized the Sega interface as an unprotectable procedure under § 102(b), saying that to get an exclusive right in that interface, Sega needed to get a patent.¹⁶² This ruling allowed Accolade (and other independent software developers) to create new non-infringing programs that could run on the popular Genesis platform and maintained the proper boundary lines between patent and copyright protections for computer programs.

4. *Because of § 102(b) Exclusions, the Scope of Copyright in Programs Is Thinner than the Scope of Copyright in Conventional Literary Works*

A fourth proposition concerning § 102(b) that should be uncontroversial is that the scope of copyright protection in computer programs is generally much thinner than the scope of copyright in conventional literary works (e.g., novels and poetry) because programs embody many functional design elements that lie outside the scope of copyright protection under § 102(b).¹⁶³ Courts have recognized that it is necessary to filter out procedures, processes, systems, and methods of operation before ruling on software copyright infringement claims.¹⁶⁴ Conventional literary works, by contrast, are typically highly expressive and non-functional, which is why courts

159. *See* 977 F.2d 1510, 1524–27 (9th Cir. 1992) (noting that the functionality of programs limits scope of copyright under § 102(b) and emphasizing that copyright should not be construed to give programmers patent-like protection for elements excluded from copyright under § 102(b)).

160. *Id.* at 1516–17. Sega did not claim that Accolade infringed because it copied the SSO of the Genesis interface.

161. *Id.* at 1526.

162. *Id.* at 1522, 1526.

163. *Comput. Assocs. Int'l, Inc. v. Altai, Inc.*, 982 F.2d 693, 712 (2d Cir. 1992).

164. *See, e.g., Gates Rubber Co. v. Bando Chem. Indus.*, 9 F.3d 823, 842–46 (10th Cir. 1993) (applying the abstraction-filtration-comparison test to the computer program).

mainly concentrate on filtration of ideas, facts, and scenes a faire elements in assessing infringement claims.¹⁶⁵

5. *SSO Obscures the Distinction between Nonliteral Elements of Programs That Are Protectable by Copyright and Those That Are Unprotectable Under § 102(b)*

A fifth proposition that should be uncontroversial is that “SSO” is not a useful way of conceptualizing which nonliteral elements of computer programs are copyright-protectable expressions that cannot be reused without infringing.¹⁶⁶ Although courts have sometimes based infringement findings on copying of the detailed structure of novels or plays,¹⁶⁷ SSO was not a term of art in the copyright field prior to the *Whelan* decision. Nor is SSO a term of art in the computing field, as the Second Circuit observed in *Altai*.¹⁶⁸ As applied to programs, SSO obscures the reality that many aspects of program structure and organization are procedures, processes, systems, and methods of operation that § 102(b) excludes from copyright’s protection.¹⁶⁹ The term fails to provide a workable framework within which courts can separate out nonliteral program expression from nonliteral functional elements excluded under § 102(b). As the District Court in *Oracle* observed, there has been a trend away from use of SSO in the post-*Altai* software copyright case law, which has been “driven by fidelity to Section 102(b)” to avoid the “danger” of overprotection of programs by copyright law.¹⁷⁰

165. This explains why Judge Hand’s patterns-of-abstraction test for infringement, which has long been applied to conventional literary and dramatic works, does not mention processes or systems. *See Nichols v. Universal Pictures*, 45 F.2d 119, 121 (2d Cir. 1930), *cert. denied*, 282 U.S. 902 (1931).

166. *See Altai*, 982 F.2d at 702–06 (questioning the *Whelan* decision’s use of SSO); *Oracle Am., Inc. v. Google Inc.*, 872 F. Supp. 2d 974, 996 (N.D. Cal. 2012), *rev’d*, 750 F.3d 1339 (Fed. Cir. 2014) (pointing out that SSO terminology had not been used in the software copyright case law since 1989); *see also Weinreb, supra* note 10, at 1170 (critical of SSO formulation for software).

167. *See, e.g., Sheldon v. Metro-Goldwyn Pictures Corp.*, 81 F.2d 49 (2d Cir. 1936).

168. *Altai*, 982 F.2d at 706 (noting that SSO reflects an inaccurate understanding of computer science).

169. Plaintiffs in software copyright cases sometimes rely on the literary work metaphor to deflect attention from the functionality of software. *See, e.g.,* Opening Brief and Addendum of Plaintiff Appellant, *Oracle of Am. v. Google Inc.*, Court of Appeals for the Federal Circuit, Case No. 13-1021, at 1–2 (likening the Android software to a knockoff of a Harry Potter novel).

170. *Oracle*, 872 F. Supp. 2d at 996.

6. Summary

These five propositions are grounded in a straightforward reading of the copyright statute and long-standing policy considerations that explain why copyright protection is “thinner” for utilitarian works than for conventional works of art and literature. Ongoing progress in the computing field requires a realm of freedom to reuse functional design elements of programs.¹⁷¹ Courts and commentators who believe that § 102(b) merely restates the idea/expression distinction should reconsider their conceptions of § 102(b).

But even accepting these propositions, the question remains whether the command structures in the *Borland* and *Oracle* cases should have been deemed unprotectable systems or methods of operation under § 102(b). To that question, we now turn.

B. *LOTUS V. BORLAND*

Lotus v. Borland is the best known of the cases addressing whether a command structure of a computer program user interface (UI) is protectable by copyright law.¹⁷² Lotus was not, however, the first software developer to sue a competitor for copyright infringement because the defendant adopted the same set of commands, organized in the same way, as the plaintiffs’ program in order to carry out the same set of program functions.¹⁷³

Courts in the early cases typically found infringement because they conceived of the plaintiffs’ command structures as compilations of words whose selection and arrangement was original enough to warrant copyright protection, even if each command word on its own (e.g., copy or paste) was unprotectable.¹⁷⁴ These cases often relied upon *Whelan*, which considered

171. See, e.g., *Altai*, 982 F.2d at 707–12 (emphasizing various elements of programs that are unprotectable by copyright law because of the need of programmers to reuse them and quoting commentary expressing concern that *Whelan* would enable programmers to lock-up basic programming techniques and give first comers quasi-monopoly power).

172. *Lotus Dev. Corp. v. Borland Int’l, Inc.*, 49 F.3d 807 (1st Cir. 1995).

173. After *Whelan*, several lawsuits were brought against makers of “clone” or “work-alike” programs that used the same or substantially similar commands as the industry leader. See, e.g., *Lotus Dev. Corp. v. Paperback Software Int’l*, 740 F. Supp. 37 (D. Mass. 1990); *Digital Comm’n Assoc., Inc. v. Softklone Distrib. Corp.*, 659 F. Supp. 449 (N.D. Ga. 1987); *Broderbund Software, Inc. v. Unison World*, 648 F. Supp. 1127 (N.D. Cal. 1986). These were sometimes known as “look and feel” cases because of similarities in the way the programs looked and the way they operated in the same or substantially similar ways. See, e.g., Pamela Samuelson, *Why the Look and Feel of Software User Interfaces Should Not Be Protected by Copyright Law*, 32 COMM. ACM 563 (May 1989). *Whelan* expressed receptivity to copyright protection for program “look and feel.” *Whelan Assoc. Inc. v. Jaslow Dental Lab., Inc.*, 797 F.2d 1222, 1231, 1245–47 (3d Cir. 1986).

174. See, e.g., *Softklone*, 659 F. Supp. at 452–59.

all structural elements of programs (e.g., UI command hierarchies) as protectable expression as long as there was more than one way to structure those elements.¹⁷⁵

Borland unquestionably copied the command hierarchy of Lotus 1-2-3 in its Quattro Pro (QP) program.¹⁷⁶ The District Court regarded Borland's appropriation of this hierarchy as infringement.¹⁷⁷ Even when Borland issued a new release of the QP emulation UI so that the Lotus command words were no longer visible, the District Court found this too infringed because QP was still utilizing the Lotus command structure, albeit invisibly.¹⁷⁸ The District Court conceived the command hierarchy to have, in effect, become a nonliteral element of the program.

Borland's reuse of the command structure of Lotus 1-2-3 was different from the earlier UI cases. Unlike those defendants, Borland had developed its own "native" UI for QP, which organized commands in a different way than Lotus 1-2-3. But Borland provided an "emulation" UI that presented users with the same commands in the same order as 1-2-3 so that prospective customers who had built macros (i.e., mini-programs for commonly executed sequences of spreadsheet functions) in the Lotus macro language could use those macros in QP.

Because the command hierarchy of Lotus 1-2-3 was "a fundamental part of the functionality of the Lotus macros,"¹⁷⁹ Borland argued it was outside the scope of copyright protection available to the Lotus program under § 102(b).¹⁸⁰ This structure, Borland believed, was akin to the structure of Selden's bookkeeping forms, and as constituent elements of Lotus' method of operation or system, it was patent, not copyright, subject matter.¹⁸¹

175. See, e.g., *id.* at 454–55.

176. *Borland*, 49 F.3d at 810.

177. *Lotus Dev. Corp. v. Borland Int'l, Inc.*, 799 F. Supp. 203, 219 (D. Mass. 1992), *rev'd*, 49 F.3d 807 (1st Cir. 1995).

178. See generally *Lotus Dev. Corp. v. Borland Int'l, Inc.*, 831 F. Supp. 223, 245 (D. Mass. 1993).

179. In *Lotus Dev. Corp. v. Paperback Software Int'l*, the District Court recognized the functionality of the Lotus macro system, which allowed users to construct mini-programs for commonly executed sequences of functions, but did not think it was relevant to copyright protection for the command hierarchy. 740 F. Supp. 37, 64 (D. Mass. 1990). Paperback had argued that macro compatibility required use of this hierarchy, but the court found this unpersuasive. *Id.* at 69.

180. Borland argued that *Altai* supported its macro-compatibility defense. *Borland*, 799 F. Supp. at 215–19.

181. Borland's argument is most clearly set forth in the First Circuit's opinion. *Lotus Dev. Corp. v. Borland Int'l, Inc.*, 49 F.3d 807, 814 (1st Cir. 1995), *aff'd by an equally divided Court*, 516 U.S. 233 (1996).

The District Court did not find the user macro interoperability argument persuasive.¹⁸² It viewed Borland's copying of the 1-2-3 command structure as infringement because the command hierarchy was not an abstract idea, but rather concrete and detailed; in that court's view, there were other ways that spreadsheet commands could be named and organized, as evidenced by QP's native UI.¹⁸³

Borland appealed its loss to the First Circuit.¹⁸⁴ The main question on appeal was what meaning to give to the exclusions set forth in § 102(b).¹⁸⁵ Given the text of the statute and the legislative history explaining Congress's reasons for including § 102(b) in the statute, it seemed unlikely that this court would rule that § 102(b) was merely a restatement of the idea/expression distinction, as the District Court had opined in *Borland*.

In its ruling in Borland's favor, the First Circuit held that the command structure was a method of operation under § 102(b), although on a different rationale than Borland argued.¹⁸⁶ The First Circuit likened the command hierarchy of Lotus 1-2-3 to the buttons for interacting with a VCR machine.¹⁸⁷ Both were, in its view, methods of operating machines to accomplish functional tasks.¹⁸⁸ As methods of operation, both were unprotectable by copyright law. Although the First Circuit invoked *Baker* in support of its holding, there was disappointingly little analysis in support of the court's conclusion, either about the implications of *Baker* or the meaning that courts should give to § 102(b).¹⁸⁹

The First Circuit viewed Borland as having literally copied the Lotus command structure, but decided that didn't matter because that structure, however creative it might be, constituted a method of operation.¹⁹⁰ It did

182. *Borland*, 799 F. Supp. at 210 (“[I]t is irrelevant that the 1-2-3 interface includes functional elements or ‘comprises a system’ so long as it includes separable expressive elements.”). In its earlier ruling against Paperback, which had made a clone of Lotus 1-2-3, the District Court dismissed as a “word game” Paperback's argument that the Lotus macro language was unprotectable by copyright law. *Paperback*, 740 F. Supp. at 72.

183. *Borland*, 799 F. Supp. at 212–13.

184. *Borland*, 49 F.3d at 812.

185. *Id.*

186. Borland argued that the facts and arguments in *Baker* were “identical” to those it was making. The First Circuit was not convinced. *Id.* at 814.

187. *Id.* at 817.

188. For a similar metaphor, see *Apple Computer v. Microsoft Corp.*, 799 F. Supp. 1006, 1023 (N.D. Cal. 1992), *aff'd*, 35 F.3d 1435 (9th Cir. 1994) (likening Apple's graphical user interface to the user interface for automobiles).

189. *Borland*, 49 F.3d at 813–14; *see also* Weinreb, *supra* note 10, at 1207 (describing the First Circuit's analysis as “too short to be satisfactory”).

190. *Borland*, 49 F.3d at 815–16.

not, however, find persuasive Borland's *Baker*-based argument that the command structure was patent, not copyright, subject matter.¹⁹¹ The First Circuit cited approvingly to *Altai*, but regarded the Second Circuit's decision as inapplicable because that case involved claims of non-literal infringement, whereas *Borland* was, in its view, a literal copying case.¹⁹² Although literal copying of content from a copyrighted work is generally more likely to be infringement, the First Circuit in *Borland* conceptualized the command hierarchy as too functional to qualify as protectable expression.¹⁹³ *Altai*, in fact, had more to teach the First Circuit on compatibility issues than the court perceived in *Borland*.¹⁹⁴

Whether the 1-2-3 command hierarchy was a literal or nonliteral element of the Lotus program is actually a more interesting question than either the District Court or the First Circuit recognized.¹⁹⁵ In some sense it was both: users of the Lotus program can, of course, see the selection and arrangement of command words when they use 1-2-3. Viewed as a compilation of words, the command hierarchy looks like a literal component of the program. Each command is, however, an abstraction that not only identifies the particular function that the command represents, but also provides users with a means to invoke that function, which, in turn, is a nonliteral element embedded in the literal text of the program. U.S. copyright law defines "computer program" as "a set of statements or instructions to be used directly or indirectly in a computer in order to bring about a certain result."¹⁹⁶ Viewed in this light, the program source and object code is the literal expression of the software, and the UI command hierarchy is an abstraction that identifies the set of functions that the

191. *Id.* at 813–14. The First Circuit did not explain why it found the argument unpersuasive. The patent or copyright subject matter issue is discussed *infra* Part V.

192. *Id.* at 814–15.

193. *Id.* at 815 ("The Lotus command hierarchy provides the means by which users control and operate Lotus 1-2-3.").

194. See Brief Amici Curiae of American Committee for Interoperable Systems and Computer & Communications Industry Ass'n in Support of Respondent, *Lotus Dev. Corp. v. Borland Int'l Inc.*, 516 U.S. 233 (1996) (No. 94-2003), 1995 WL 728487 (explaining the relevance of *Altai* in the *Borland* case). Later cases, relying in part on *Altai*, recognized that even literal copying may be excusable when necessary for achieving interoperability. See, e.g., *Bateman v. Mnemonics, Inc.*, 79 F.3d 1532, 1547 (11th Cir. 1996).

195. Weinreb noted the oddity of the parties' positions about the nature of programs that copyright could protect: Lotus taking a broad conception of programs, wanting to tie its UI closely to the code, and Borland taking a narrow view, as though the code constituted the program and the UI was a result of the program. Weinreb, *supra* note 10, at 1154–63.

196. 17 U.S.C. § 101 (definition of "computer program").

program is capable of executing. The UI command structure is, moreover, among the results that the program code produces.

The nonliteral character of the command hierarchy should have been evident when Borland introduced the key reader version of QP, which allowed macros constructed in 1-2-3 to be executed, even though users could no longer see the commands.¹⁹⁷ The First Circuit should have recognized the macro system as a nonliteral element of the Lotus program because it was neither visible to a 1-2-3 user through the UI, nor was it associated with particular blocks of code.¹⁹⁸

When the Supreme Court decided to hear Lotus's appeal, it seemed that the Court would finally provide an answer to the long-simmering question of how courts should interpret the method and system exclusions of § 102(b) as applied to computer programs. However, the Court's 4-4 split affirmed the First Circuit ruling without setting a precedent or resolving the circuit split on this issue.¹⁹⁹

The deep split within the Court may have been due to some Justices agreeing with the literal infringement approach taken by the District Court, while other Justices may have seen some merit in the First Circuit's interpretation of § 102(b), as cryptic as that court's analysis was, or in Borland's argument, that the command hierarchy was patent, not copyright, subject matter.²⁰⁰

The sounder analysis supporting the ruling in Borland's favor would have focused on the essential role that the Lotus command structure played in facilitating the functionality of the Lotus macro system. As the District Court recognized, the command hierarchy was a "fundamental part of the

197. *Lotus Dev. Corp. v. Borland Int'l, Inc.*, 831 F. Supp. 223, 228 (D. Mass. 1993) ("[T]he Key Reader file contains a virtually identical copy of the Lotus menu tree structure, but represented in a different form and with first letters of menu command names in place of the full menu command names."). The District Court found infringement of the key reader version of QP. *Id.* at 231. This, in effect, extended copyright protection to the functional behavior of the program. See Pamela Samuelson, *Brief Amicus Curiae of Copyright Law Professors* in *Lotus Development Corp. v. Borland Int'l, Inc.* (brief to U.S. Supreme Court), 3 J. INTELL. PROP. L. 103, 131-32 (1996) (explaining why the District Court's ruling would have extended protection to the functional behavior of programs and why copyright should not extend so far).

198. An analogy may help to clarify the relationship among these components: The UI of a program is akin to the face of a clock. The program code is like the mechanism inside the clock that causes the hands of the clock to move. The macro system is like a specialized part of the clock (e.g., the component that users can set to cause an alarm to ring at a particular time).

199. *Lotus Dev. Corp. v. Borland Int'l, Inc.*, 516 U.S. 233 (1995).

200. Brief for Respondent at 22-37, *Lotus Dev. Corp. v. Borland Int'l Inc.*, 516 U.S. 233 (1996) (No. 94-2003), 1995 WL 728538.

functionality of the Lotus macros.”²⁰¹ That is, the Lotus command structure was a critical component in the functioning of the macros because user-created macros would not execute in QP unless exactly the same commands were available, arranged in exactly the same order.²⁰² The command structure was thus a constituent element of the macros system that should have been outside the scope of the Lotus copyright under § 102(b).²⁰³ Alternatively, the Court could have recognized that, in keeping with *Altai*, there were external factors (i.e., the investment of users who had constructed macros in the Lotus macro language) that constrained the design choices of second comers (i.e., Borland because the command structure was essential to achieving compatibility with user macros).²⁰⁴

Judge Boudin’s concurring opinion in *Borland* suggests the macro compatibility consideration resonated with him:

[I]t is very hard to see that Borland has shown any interest in the Lotus menu except as a fall-back option for those users already committed to it by prior experience or in order to run their own macros using 1-2-3 commands . . . [I]t is unlikely that users who value the Lotus menu for its own sake—independent of any investment they have made themselves in learning Lotus’

201. *Lotus Dev. Corp. v. Paperback Software Int’l*, 740 F. Supp. 37, 64 (D. Mass. 1990).

202. See Pamela Samuelson, *Computer Programs, User Interfaces, and Section 102(b) of the Copyright Act of 1976: A Critique of Lotus v. Paperback*, 55 LAW & CONTEMP. PROB. 311, 331–37 (1992) (explaining why the Lotus macro system was unprotectable under § 102(b)); see also Brief Amicus Curiae of Copyright Professors in Support of Respondent at 3–5, *Lotus Dev. Corp. v. Borland Int’l Inc.*, 516 U.S. 233 (1996) (No. 94-2003), 1995 WL 728563 (setting forth alternative theories about the application of § 102(b) to the Lotus command hierarchy).

203. That methods and systems and their constituent parts are unprotectable by copyright law was recently affirmed in the Ninth Circuit’s *Bikram’s Yoga* decision. See *Bikram’s Yoga*, 803 F.3d at 1039 (“An essential element of this ‘system’ is the order in which the yoga poses and breathing exercises are arranged.”); *id.* at 1042 (“[T]he medical and functional considerations at the heart of the [Bikram] Sequence compel the very selection and arrangement of poses and breathing exercises for which he claims copyright protection.”).

204. See Brief Amici Curiae of American Committee for Interoperable Systems and Computer & Communications Industry Ass’n in Support of Respondent, *Lotus Dev. Corp. v. Borland Int’l Inc.*, 516 U.S. 233 (1996) (No. 94-2003), 1995 WL 728487 (explaining the relevance of *Altai* in the *Borland* case). A second alternative analysis would have framed *Borland* as an instance in which functionality and expression had, in effect, merged because of the role of the Lotus command hierarchy in enabling the functionality of the macro system. See, e.g., Brief of the United States as Amicus Curiae at 19–21, *Google Inc. v. Oracle Am., Inc.*, 135 S. Ct. 2887 (2015) (No. 14-410), 2015 WL 2457656 (suggesting a merger rationale for the *Borland* ruling); see also *infra* text accompanying note 382 for a discussion of merger in relation to *Borland*.

commands or creating macros dependent upon them—would choose the Borland program in order to secure access to the Lotus menu.²⁰⁵

Judge Boudin also took note of the lock-in effects that would result from a ruling in Lotus' favor:

[I]t is hard to see why customers who have learned the Lotus menu and devised macros for it should remain captives of Lotus because of an investment in learning made by the users and not by Lotus. Lotus has already reaped a substantial reward for being first; assuming that the Borland program is now better, good reasons exist for freeing it to attract old Lotus customers: to enable the old customers to take advantage of a new advance, and to reward Borland in turn for making a better product. If Borland has not made a better product, then customers will remain with Lotus anyway.²⁰⁶

It is unfortunate that Judge Boudin did not find a way to connect this concern to the Congressional intent that § 102(b) serve as a statutory tool through which courts could take competition and ongoing innovation policy considerations into account in construing the scope of copyright protection in programs.²⁰⁷ He considered the majority's interpretation to be "defensible," although he thought that fair use would be a closer doctrinal fit, suggesting there was need for a new doctrine to address such considerations.²⁰⁸

A compatibility-based § 102(b) argument should have prevailed in *Borland*. Subsequent cases have done a somewhat better job than the First Circuit in employing § 102(b) in software copyright cases.²⁰⁹ The Ninth Circuit reaffirmed in 2000 that program interfaces necessary for achieving interoperability are unprotectable procedures under § 102(b),²¹⁰ and has

205. *Lotus Dev. Corp. v. Borland Int'l, Inc.*, 49 F.3d 807, 820 (1st Cir. 1995), *aff'd by an equally divided Court*, 516 U.S. 233 (1996) (Boudin, J., concurring).

206. *Id.* at 821; *see also* Dan L. Burk, *Method and Madness in Copyright Law*, 2007 UTAH L. REV. 587, 591–92 (2007).

207. *See supra* note 4 and accompanying text.

208. *Borland*, 49 F.3d at 821–22 (Boudin, J., concurring).

209. *See, e.g.*, *Incredible Techs. v. Virtual Techs., Inc.*, 284 F. Supp. 2d 1069, 1078 (N.D. Ill. 2003), *aff'd*, 400 F.3d 1007, 1012 (7th Cir. 2005); *Mitek Holdings, Inc. v. Arce Eng'g Co.*, 89 F.3d 1548, 1556–57 (11th Cir. 1996); *Ilog, Inc. v. Bell Logic LLC*, 181 F. Supp. 2d 3, 12–14 (D. Mass. 2002); *Torah Software Ltd. v. Drosnin*, 136 F. Supp. 2d 276, 291–92 (S.D.N.Y. 2001); *O.P. Solutions v. Intell. Prop. Network Ltd.*, No. 96 Civ. 7952 (LAP), 1999 WL 47191, at *16–18 (S.D.N.Y. Feb. 2, 1999).

210. *Sony Computer Entm't, Inc. v. Connectix Corp.*, 203 F.3d 596, 602–03 (9th Cir. 2000).

also held that a list of commands for the user interface of a computer program is unprotectable by copyright law.²¹¹

It is fair to say that the First Circuit could have done a better job explaining why the Lotus command hierarchy was an unprotectable system or method of operation under § 102(b). This Section has explained that this ruling was sound and consistent with the competition and innovation policies expressed in other software copyright decisions that have construed § 102(b) as a limiting principle on the scope of copyright when compatibility issues are at stake.

C. *ORACLE V. GOOGLE*

Google's confidence that it could lawfully use command structures derived from the Java API was built partly on the *Borland* decision's interpretation of § 102(b) and partly on the Ninth Circuit's ruling that the Sega interface was unprotectable under § 102(b).²¹² In the twenty years since the Supreme Court punted on interpreting § 102(b) in *Borland*, the First Circuit's decision has been met with mostly positive reactions in the

211. *Ashton-Tate Corp. v. Ross*, 916 F.2d 516, 521–22 (9th Cir. 1990) (rejecting Ross's claim of joint authorship based on having provided a list of commands for a computer program on which he and another programmer were working, and agreeing with the District Court that the list of commands was unprotectable under § 102(b)).

212. See Brief of Appellee and Cross-Appellant Google Inc., Oracle Am., Inc. v. Google Inc., No. 2013-1021 at 33–52 (relying on *Sega v. Accolade*), 57–65 (relying on *Lotus v. Borland*) (Fed. Cir. 2013). Oracle sued Google in 2010 for infringing patents that Oracle claimed read onto component elements of the Android mobile phone platform and for infringing Oracle's copyright in the Java platform, which was said to include code, specifications, documentation and other materials. *Oracle Am., Inc. v. Google Inc.*, 750 F.3d 1339, 1347–48 (Fed. Cir. 2014). The jury ruled against Oracle's patent claims and against its claims of copyright infringement in Java documentation; however, the jury found Google to have infringed copyright in a range by checking the subroutine and in structure of certain Java API packages (assuming that these were copyright-protectable, an issue which the District Court reserved for itself). *Id.* Oracle's appeal of the copyright ruling went to the CAFC because of the patent claim in the complaint. The CAFC acknowledged that it was obliged to follow Ninth Circuit precedent. *Id.* at 1353. This section will show that it did not do so.

courts.²¹³ Many have followed it,²¹⁴ some have distinguished it,²¹⁵ and a few have criticized or rejected it.²¹⁶ The trial court in *Oracle* was one of the followers.²¹⁷ The CAFC, on the other hand, both distinguished and criticized *Borland* in its decision in favor of Oracle.²¹⁸

Although the facts of the *Oracle* case are more complicated than those in *Borland*, the legal question in the two cases is remarkably similar: whether a command structure designed to be implemented in computer program code is unprotectable under § 102(b). In *Oracle*, the relevant

213. *Borland* had been cited 133 times as of July 6, 2015, in the Lexis database. Of these, 115 were positive citations; only 6 either distinguished or criticized the ruling. *Borland* has been cited at least once in every circuit.

214. See, e.g., *Hutchins v. Zoll Med. Corp.*, 492 F.3d 1377, 1383–85 (Fed. Cir. 2007); *Incredible Technologies*, 400 F.3d at 1012; *Mitek*, 89 F.3d at 1556–57; *Wyatt Tech. Corp. v. Malvern Instr. Inc.*, 2009 U.S. Dist. LEXIS 66097, at *6 (C.D. Cal. July 29, 2009); *Jamison Bus. Sys., Inc. v. Unique Software Support Corp.*, 2005 WL 1262095, at *12–13 (E.D.N.Y. May 26, 2005); see also *Lexmark Int’l, Inc. v. Static Control Components*, 387 F.3d 522, 538 (6th Cir. 2004) (citing *Borland* approvingly); *O.P. Solutions*, 1999 WL 47191, at *8 (citing *Borland* approvingly).

215. See, e.g., *Maddog Software, Inc. v. Sklader*, 382 F. Supp. 2d 268, 278–82 (D. N.H. 2005) (applying *Altai* filtration in software copyright case).

216. See, e.g., *Mitel, Inc. v. Iqtel, Inc.*, 124 F.3d 1366, 1371–72 (10th Cir. 1997) (rejecting § 102(b) defense based on *Borland*, but affirming non-infringement ruling because command codes were either unoriginal or unprotectable as scenes a faire). Most commentators, however, have cited to *Borland* approvingly. See, e.g., Christina Bohannon, *Reclaiming Copyright*, 23 CARDOZO ARTS & ENT. L.J. 567, 592–93 (2006); Burk, *supra* note 206, at 591–92; Thomas F. Cotter, *The Procompetitive Interest in Intellectual Property Law*, 48 WM. & MARY L. REV. 483, 510 n.115 (2006); Stacey L. Dogan & Joseph P. Liu, *Copyright Law and Subject Matter Specificity: The Case of Computer Software*, 61 N.Y.U. ANN. SURV. AM. L. 203, 211–12 (2005); Herbert Hovenkamp, *Response: Markets in IP and Antitrust*, 100 GEO. L.J. 2133, 2144 n.54 (2012); Dennis S. Karjala, *A Coherent Theory for the Copyright Protection of Computer Software and Recent Judicial Interpretations*, 66 U. CIN. L. REV. 53, 105–07 (1997); Peter Lee, *The Evolution of Intellectual Infrastructure*, 83 WASH. L. REV. 39, 84–85 (2008); Aaron K. Perzanowski, *Rethinking Anticircumvention’s Interoperability Policy*, 42 U.C. DAVIS L. REV. 1549, 1563 n.39 (2009); Michael Risch, *How Can Whelan v. Jaslow and Lotus v. Borland Both Be Right? Reexamining the Economics of Computer Software Reuse*, J. MARSHALL J. COMPUTER & INFO. L. 511, 545–46 (1999).

217. *Oracle Am., Inc. v. Google Inc.*, 872 F. Supp. 2d 974, 990–91 (N.D. Cal. 2012).

218. *Oracle Am., Inc. v. Google Inc.*, 750 F.3d 1339, 1364–68 (Fed. Cir. 2014). The CAFC distinguished *Borland* because that defendant had not literally copied any Lotus code, whereas in the CAFC’s view, Google had literally infringed Oracle code. *Id.* at 1365. However, it disagreed with *Borland*’s conclusion that creative methods of operation were ineligible for copyright protection. *Id.* at 1366–67. The CAFC’s decision is, in this respect, at odds with Ninth Circuit precedents. See, e.g., *Bikram’s Yoga Coll. of India, L.P. v. Evolution Yoga, LLC*, 803 F.3d 1032, 1035–44 (9th Cir. 2015) (recognizing that Bikram’s selection and arrangement of yoga poses and breathing exercises was creative, but rejecting its copyright claims because the poses and exercises were constituent elements of an unprotectable method or system).

command structure drew upon 37 of the 166 Java API packages that Google incorporated into Android.

Some technical background is necessary to inform the legal analysis. The Java API is tightly organized in a set of 166 “packages.”²¹⁹ Each package identifies types of functions which specific elements of the Java API make available to programmers.²²⁰ The engineers who developed the Java API assigned a specific name to each package. Within each package are numerous related classes of functions. Each class consists of numerous variables and method headers (sometimes called declarations) that specify the type of subroutine to be carried out when the command for that function is invoked.²²¹ The 37 Java API packages at issue in *Oracle* consisted of 600+ classes, and 6000+ method headers. Google believed that these classes and method headers were unprotectable elements of the Java API,

219. The District Court made extensive findings of fact about the Java API packages at issue. See *Oracle*, 872 F. Supp. 2d at 977–83. It indicated that all of the “declarative fact statements set forth in the order are factual findings.” *Id.* at 977 n.3.

220. For example, `java.math` is the package for programming arithmetic functions, whereas `java.awt.font` is the package for classes and interfaces for fonts. For short descriptions of the Java API packages, see <https://docs.oracle.com/javase/1.5.0/docs/api/overview-summary.html> [<https://perma.cc/EU3L-39HA>].

221. For short descriptions of the functionally related sets of classes and declarations within packages, see <http://docs.oracle.com/javase/1.5.0/docs/guide/> [<https://perma.cc/T9LC-G8R4>].

The command structure of the Java API has a standardized format: package.Class.method. An example method header is:

```
package java.lang;  
  
public class Math {  
    public static int max(int x, int y);  
}
```

This specification identifies the Java command `java.lang.Math.max`, which calls for carrying out the mathematical operation of comparing two numbers and returning the larger of the two. The following line of code uses (or “invokes”) that method for computing the maximum of the values stored in variables `i` and `j`, storing the resulting value in variable `k`:

```
k = java.lang.Math.max(i, j);
```

This command may be abbreviated:

```
k = Math.max(i, j);
```

or even further abbreviated as: `max()`.

contending, as Borland had before it, that its reimplementations of these unprotectable elements in independently written code did not infringe.²²²

The District Court recognized that “[t]he overall name tree [of the Java API packages] has creative elements,” but it was also “a precise command structure—a utilitarian and functional set of symbols, each to carry out a preassigned function.”²²³ Because of this, the court concluded that “[t]his command structure is a system or method of operation under Section 102(b) of the Copyright Act and, therefore, cannot be copyrighted.”²²⁴ It perceived its ruling to be consistent with the First Circuit’s decision that the command structure of Lotus 1-2-3 was uncopyrightable under § 102(b).²²⁵

The District Court made a finding that compatibility considerations explained why Google had drawn upon the Java API for the Android platform: “Google believed Java application programmers would want to find the same 37 sets of functionalities in the new Android system callable by the same names as used in Java. Code already written in the Java language would, to this extent, run on Android and thus achieve a degree of interoperability.”²²⁶ Java programmers, the court noted, had written millions of lines of code using these method headers and classes, and reuse of that code on the Android platform required use of the same method headers.²²⁷ Although Google copied the exact names and functions of the Java API classes and method headers, the court noted that Google “took care to use different code to implement the six thousand-plus subroutines (methods) and six-hundred-plus classes.”²²⁸

Oracle’s appeal found a very receptive audience in the CAFC. There were three main bases for the CAFC’s decision. One concerned the “copyrightability” issue in the case. The CAFC agreed with Oracle that the API packages at issue were protectable by copyright law under § 102(a)

222. *Oracle*, 872 F. Supp. 2d at 998.

223. *Id.* at 976–77.

224. *Id.* at 977.

225. *Id.* at 990–91. While concluding that § 102(b) provided a sound basis for its ruling in Google’s favor, the District Court proffered the merger doctrine as an alternative ground because “[u]nder the rules of Java, [the method headers] *must be identical* to declare a method specifying the *same* functionality.” *Id.* at 976. The merger doctrine is discussed in Part IV.

226. *Id.* at 978. The District Court also held that the names of the individual commands were unprotectable under the copyright doctrine that words and short phrases cannot be copyrighted. *Id.* at 997. Although I will not address this aspect of the District Court’s ruling, it is not implausible given judicial receptivity to this defense in rulings such as *Southco, Inc. v. Kanebridge Corp.* See 390 F.3d 276, 285 (3d Cir. 2004).

227. *Oracle*, 872 F. Supp. 2d at 1000.

228. *Id.* at 977.

because Sun's engineers had been highly creative in designing the Java API, including in naming and organizing the method headers and the classes of the API packages.²²⁹ Hence, copyright's originality standard was easily met and the resulting expression was copyright-protectable. The CAFC thought that the protection conferred by § 102(a) could not be taken away by § 102(b).²³⁰ As in *Whelan*, the CAFC perceived § 102(b) to be a restatement of the idea/expression distinction, which had no pertinence in a case involving a highly detailed structure, such as the Java API packages.²³¹

However, the CAFC did not correctly understand the copyrightability issue in the case. Courts use that term in three different ways: sometimes they use it to indicate that a particular work is (or is not) proper subject matter for copyright protection (e.g., a literary work),²³² sometimes to indicate that the originality and fixation requirements have (or have not) been met as to a particular work,²³³ and sometimes to indicate that the aspect of a work that the defendant copied was outside the scope of copyright protection available to that work.²³⁴

The CAFC seems to have regarded the District Court as having used the term in the first or second sense, which explains why it regarded Google's § 102(b) defense as calling into question the copyrightability of all program code.²³⁵ However, the District Court in *Oracle*, like the First Circuit in *Borland*, used the term "copyrightability" in the third sense of the term. The District Court did not question the copyrightability of the work at issue in the case, namely, the Java Special Edition 5.0 document from which Google drew the 37 API packages.²³⁶ Rather, the District Court understood Oracle

229. *Oracle Am., Inc. v. Google Inc.*, 750 F.3d 1339, 1356 (Fed. Cir. 2014).

230. *Id.* at 1556–57.

231. *Id.* at 1367–68.

232. *See, e.g., Kelley v. City of Chicago*, 635 F.3d 290, 306 (7th Cir. 2011) (gardens not copyrightable subject matter).

233. *See, e.g., Feist Publ'ns, Inc. v. Rural Tel. Serv. Co.*, 499 U.S. 340, 362–63 (white pages listings of telephone directories are uncopyrightable because they lack sufficient originality to support copyright).

234. *Lotus Dev. Corp. v. Borland Int'l, Inc.*, 49 F.3d 807, 815 (1st Cir. 1995), *aff'd by an equally divided Court*, 516 U.S. 233 (1996) (considering whether the Lotus command hierarchy was "copyrightable" part of the Lotus program).

235. *Oracle Am., Inc. v. Google Inc.*, 750 F.3d 1339, 1379–81 (Fed. Cir. 2014).

236. The work of authorship at issue in *Oracle* was not, in fact, a computer program, as the CAFC opinion seems to imply, *id.* at 1355–56, but was a textual document entitled Java Special Edition 5.0, which sets forth the component parts of the Java API. *Oracle Am., Inc. v. Google Inc.*, 872 F. Supp. 2d 974, 979 (N.D. Cal. 2012). The Java Platform Special Edition document sets forth the specification of the Java API as well as providing other pertinent Java platform information. The current version of this document can be found at

to be claiming that Google’s use of the 6000+ method headers constituted a non-literal infringement of that document.²³⁷ The District Court understood the *Oracle* case to present a scope of protection issue, that is, whether the command structure of these method headers was protectable by copyright law.²³⁸

Second, the CAFC accepted Oracle’s claim on appeal that Google literally copied 7000 lines of Oracle source code, as well as nonliterally copied the SSO of the Java API classes.²³⁹ However, this directly contradicted the District Court’s factual finding that there was no literal copying of program code. “[A]ll agreed,” it said, “that Google had not literally copied the software but had instead come up with its own implementations of the 37 API packages.”²⁴⁰ The CAFC did not understand that the API elements that Google used in Android were not in themselves software; they were specifications that identified functions that the program was designed to perform. In order to write programs to implement these functions in accordance with rules set forth in the Java API, the source code must include the API method headers (or declarations); those method headers, however, do not become part of the executable code.²⁴¹ The CAFC’s misperception about the literal or nonliteral copying issue reinforced its conception that Google’s defense would undermine software copyright protection.

Third, the CAFC disagreed strongly with the District Court about Google’s compatibility defense. The CAFC categorically denied that the case law had recognized a “compatibility exception” to copyrightability of

<https://docs.oracle.com/javase/8/> [<https://perma.cc/4AUB-JAVN>]. This document can be downloaded for free from a public website.

237. *Id.* at 975 (all agreed that Google had not literally copied Oracle software).

238. *Id.* at 975–76.

239. *Oracle*, 750 F.3d at 1361. Oracle should not have been able to change its theory of the facts on appeal to make Google’s defense seem weaker. Had the Supreme Court taken the case, *Oracle* should have been treated as a nonliteral infringement case, not a literal copying case.

240. *Oracle*, 872 F. Supp. 2d at 975. Under the Supreme Court’s recent decision in *Teva Pharm. v. Sandoz*, 574 U.S. ___, 135 S. Ct. 831 (2015), the CAFC should have deferred to lower court findings of fact and reversed them only when the findings were clearly in error.

241. Like the District Court in *Oracle*, one post-*Oracle* decision has distinguished declaring code (i.e., method headers), the function of which was to call for the performance of a particular function, and implementing code, the function of which was to be the object code to carry out that function. *SAS Inst., Inc. v. World Programming Ltd.*, 64 F. Supp. 3d 755, 777 (E.D.N.C. 2014).

program APIs.²⁴² Echoing *Franklin*, the CAFC characterized compatibility “as a commercial and competitive objective,” which had no bearing on copyrightability.²⁴³ In addition, it perceived no evidence that the design of the Java API packages were constrained by “compatibility requirements” of pre-existing programs, the only constraints it thought might have significance.²⁴⁴ The CAFC was also skeptical of the defense, saying “Google designed Android so that it would *not* be compatible with the Java platform,”²⁴⁵ thereby defeating the Java goal of allowing programmers to “write once, [code that will] run everywhere.”²⁴⁶ This frontal attack, not only on Google’s compatibility defense in *Oracle*, but on compatibility defenses in software copyright cases, requires a more elaborate response.

D. THE IMPLICATIONS OF § 102(b) FOR COMPATIBILITY DEFENSES

The functionality of computer program APIs is unquestionable insofar as programs can interoperate with existing programs only if they conform to the interface procedures that existing programs have adopted. This explains why the Ninth Circuit has twice unequivocally affirmed that APIs, insofar as they constitute the functional requirements for achieving compatibility with other programs, are unprotectable procedures under § 102(b).²⁴⁷ The Ninth Circuit has thus recognized in these and subsequent cases that § 102(b) is more than merely a restatement of the idea/expression distinction,²⁴⁸ a proposition that the CAFC refused to accept in its *Oracle* decision.²⁴⁹ For more than two decades, courts have consistently followed

242. *Oracle*, 750 F.3d at 1368–71. Yet, in *Atari Games*, the CAFC intimated that no infringement might have been found if AG had copied only what was necessary for interoperability. 975 F.2d 832, 844–45 (Fed. Cir. 1992).

243. *Oracle*, 750 F.3d at 1371 (quoting *Apple Comput., Inc. v. Franklin Comput. Corp.*, 714 F.2d 1240, 1253 (3d Cir. 1983)).

244. *Id.*

245. *Id.*

246. *Id.* at 1348.

247. See *Sony Comput. Entm’t, Inc. v. Connectix Corp.*, 203 F.3d 596, 602–05 (9th Cir. 2000); *Sega Enters. Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1522 (9th Cir. 1992).

248. See also *Bikram’s Yoga*, 803 F.3d at 1041 (“That the sequence may possess many constituent parts does not transform it into a proper subject matter of copyright protection. Virtually any process or system could be dissected in a similar fashion.”). Defendants in the *Oracle*, *Cisco v. Arista* and *Synopsys v. ATopTech* cases tried to persuade the Ninth Circuit to amend its *Bikram’s Yoga* decision to clarify that the CAFC has misunderstood Ninth Circuit law concerning § 102(b). See Motion for Leave to File Brief Amici Curiae of Google Inc., Arista Networks, & ATopTech, Inc., and Brief Amicus Curiae of Google Inc., Arista Networks, & ATopTech, Inc., at 3–12, *Bikram’s Yoga Coll. of India, L.P. v. Evolution Yoga, LLC*, 803 F.3d 1032 (2015) (No. 13-55763).

249. See *Oracle*, 750 F.3d at 1365–67. The CAFC has previously recognized that § 102(b) has broader significance. See *Hutchins v. Zoll Med. Corp.*, 492 F.3d 1377, 1383

the rulings in *Sega* and *Altai* that program interfaces necessary for interoperability are unprotectable by copyright law,²⁵⁰ making the CAFC's *Oracle* decision an outlier. Although the CAFC recognized in *Oracle* that it should apply Ninth Circuit case law,²⁵¹ and even purported to do so,²⁵² its ruling is at odds with Ninth Circuit precedents. Because the Supreme Court decided not to review the CAFC ruling,²⁵³ software developers now face an uncertain future when they raise compatibility defenses.²⁵⁴

The CAFC's *Oracle* decision is most worrisome in its outright repudiation of software compatibility defenses, its resurrection of the anti-compatibility *Franklin* dictum, and its refusal to acknowledge that the Ninth Circuit in *Sega* and *Connectix* treated program interfaces as unprotectable procedures under § 102(b). As amicus curiae briefs of computer scientists, software companies, industry associations, and public interest groups filed in support of Google's petition for certiorari attest,²⁵⁵ freedom to reuse APIs, insofar as they are necessary for interoperability, promotes healthy competition and ongoing innovation in the software industry. More than two decades of copyright rulings have endorsed this freedom, and the software industry has flourished under this legal regime.²⁵⁶ While one can

(Fed. Cir. 2007) (“[C]opyright protection does not extend to the methods that are performed with program guidance.”).

250. See *infra* text accompanying notes 259–64 for a review of those cases.

251. *Oracle*, 750 F.3d at 1353. *Oracle*'s appeal of the District Court's copyright ruling went to the CAFC, instead of the Ninth Circuit, because *Oracle* had alleged patent, as well as copyright, infringement. *Id.*

252. See 2 PAUL GOLDSTEIN, GOLDSTEIN ON COPYRIGHT, § 2.15.2.1 (3d ed. 2005 & Supp. 2016) (arguing CAFC purported to follow Ninth Circuit law in *Oracle*). The CAFC disingenuously interpreted *Sega* as having considered compatibility as a factor in fair use analysis, choosing to ignore the court's unequivocal statements about interface procedures being excluded from copyright under § 102(b). See *Oracle*, 750 F.3d at 1369–70. The CAFC said that compatibility could be considered in connection with Google's fair use defense on remand. *Id.* at 1372–77. In rejecting the District Court's copyrightability ruling, the CAFC relied on its decision in *Atari Games Corp. v. Nintendo of Am., Inc.*, 897 F.2d 1572 (Fed. Cir. 1990) (also rejecting a software compatibility defense), which the CAFC regards as a correct interpretation of Ninth Circuit law. *Oracle*, 750 F.3d at 1353–61.

253. *Google Inc. v. Oracle Am.*, 135 S. Ct. 1021 (2015).

254. At risk are compatibility defenses in the *Cisco v. Arista* and *Synopsys v. AT&T* cases because the plaintiffs included a patent claim in the case, meaning any appeal will go to the CAFC. See *supra* note 30.

255. Briefs available online at *Google Inc. vs. Oracle America, Inc.*, SCOTUSBLOG, <http://www.scotusblog.com/case-files/cases/google-inc-v-oracle-america-inc/> [<https://perma.cc/XC5H-78ZL>].

256. See, e.g., Brief Amicus Curiae of Computer and Communications Industry Ass'n in Support of Petitioner at 4–8, *Oracle*, 135 S. Ct. 1021 (2015) (No. 14-410), 2014 WL 5868946 (emphasizing importance of pro-compatibility decisions in fostering competition and innovation in the software industry).

hope that courts in the future will reject the CAFC's analysis in *Oracle*, or distinguish *Oracle* from other compatibility cases,²⁵⁷ that decision has reopened a longstanding debate about the scope of copyright in computer programs, about the implications of § 102(b), as well as about compatibility defenses. The CAFC would shunt compatibility considerations away from § 102(b) and merger defenses, leaving them to the case-by-case vagaries of fair use.²⁵⁸

Before attempting to grapple with the different perspectives in the District Court and CAFC's interpretations of Google's compatibility and § 102(b) defenses, it is useful to recognize that several different types of interoperability issues have arisen in prior cases.

1. Competing Applications' Compatibility with the Same Operating Systems: In *Altai*, the litigants were competitors in the market for scheduling programs designed to interoperate with IBM operating system programs.²⁵⁹ Similarities in the parameter lists for these programs were largely due to the need to conform to IBM's interface procedures.

2. Unlicensed Application Developer Compatibility with a Popular Platform: In *Sega*, Accolade wanted to adapt an existing videogame program so it would run on the Sega Genesis platform.²⁶⁰

3. Emulation Software to Enable Applications Developed for a Popular Platform to Run on an Alternative Platform: Connectix developed a program to emulate the functionality of the Sony PlayStation

257. The only post-*Oracle* decision to address the substance of the CAFC's ruling so far was *SAS Inst., Inc. v. World Programming Ltd.*, 64 F. Supp. 3d 755, 777 (E.D.N.C. 2014). SAS argued that the CAFC's ruling undermined WPL's compatibility defense. WPL developed software that emulated the functionality of the SAS statistical analysis program so that users of the SAS program could switch to its program and still reuse the scripts (mini-programs) they had constructed in the SAS language. The court distinguished the CAFC's ruling, saying that WPL had only used the SAS language, which under the *Oracle* decision, all were free to use. *Id.* at 776–78.

258. See *Oracle*, 750 F.3d at 1371–77. The CAFC remanded the case for retrial on the fair use issue. *Id.* at 1376–77. See *infra* notes 363, 455 and accompanying text for a discussion of compatibility issues in the context of fair use.

259. *Comput. Assocs. Int'l, Inc. v. Altai, Inc.*, 982 F.2d 693, 714–15 (2d Cir. 1992).

260. *Sega Enters. Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1514–15 (9th Cir. 1992). Similar to *Sega* was *Atari Games Corp. v. Nintendo of Am., Inc.*, 975 F.2d 832 (Fed. Cir. 1992). Like Accolade, Atari Games (AG) sought to develop videogames that would run on a popular platform. *Id.* at 836–37. The CAFC rejected AG's interoperability defense because AG had copied more from Nintendo's program than was necessary to achieve interoperability. *Id.* at 845.

platform, so that games developed for that platform could be played on the defendant's virtual machine.²⁶¹

4. Emulation of a Feature to Enable Prospective Customers' Mini-Programs Created in One Program to Run on a Competing Program: Borland created an emulation user interface to enable interoperability with user-created macros.²⁶²

5. Development of a New Operating System to Enable Continued Use of an Application Program: In *Bateman v. Mnemonics, Inc.*, the defendants had to develop their own operating system to interoperate with an application program they had developed for their business before Bateman terminated their license to use his OS to run the application.²⁶³

6. Reuse of Software to Achieve Compatibility with Hardware: In *Lexmark*, Static Control developed chips loaded with a copy of a Lexmark program so that Lexmark's competitors could make printer cartridges that would successfully interoperate with Lexmark printers.²⁶⁴

Despite the factual and contextual differences among these cases, the courts in each held that the interfaces necessary for achieving technical interoperability were unprotectable elements of copyrighted software. The CAFC in *Oracle*, however, refused to acknowledge this.²⁶⁵ These six types

261. *Sony Comput. Entm't, Inc. v. Connectix Corp.*, 203 F.3d 596 (9th Cir. 2000). *Connectix* was like *Franklin*, except that Connectix reimplemented the interface in non-infringing code instead of copying the platform code bit for bit, as Franklin had. *Id.* at 601.

262. *Lotus Dev. Corp. v. Borland Int'l, Inc.*, 49 F.3d 807, 810 (1st Cir. 1995). Similar to *Borland* was *SAS Inst., Inc. v. World Programming Ltd.*, 2013 EWCA Civ. 1482 (UK Ct. Ap. 2013) (not infringement to emulate the interface and functionality of SAS software to enable users to port over to another platform the mini-programs they had constructed in the SAS language). *See also SAS v. WPL*, 64 F. Supp. 3d at 775–78; *Bay State Tech., Inc. v. Bentley Sys., Inc.*, 946 F. Supp. 1079, 1088 (D. Mass. 1996) (not infringement for CAD software competitor to develop data translator software to allow files created in plaintiff's program to be "read" in the defendant's program).

263. *Bateman v. Mnemonics, Inc.*, 79 F.3d 1532, 1536–40 (11th Cir. 1996). The defendants reverse-engineered Bateman's OS to discern elements needed to achieve compatibility, which they claimed required copying of some literal code. *Id.* at 1539 n.18, 1545.

264. *Lexmark Int'l, Inc. v. Static Control Components*, 387 F.3d 522, 530–31 (6th Cir. 2004); *see also Secure Services*, 722 F. Supp. at 1356–64 (not infringement to reimplement competitor's variation of a standard protocol so new entrant could sell interoperable secure fax machines to the U.S. government and its contractors); *NEC Corp. v. Intel Corp.*, 10 U.S.P.Q.2d 1177, 1188 (N.D. Cal. 1989) (rejecting Intel's claim of infringement of microcode that was necessary to enable NEC's hardware to be compatible with Intel's hardware). *But see Compaq v. Procomm*, 908 F. Supp. 1409 (S.D. Tex. 1995) (parameters and their sequence not protectable, but preliminary injunction issued because defendant copied more than was necessary to achieve compatibility with Compaq hard drives).

265. *Oracle*, 750 F.3d at 1370–71.

of technical compatibility cases might seem to suggest that APIs, by their nature, should be considered unprotectable elements of computer programs. The District Court in *Oracle* did not interpret these precedents so broadly; indeed, it rejected Google's legal argument to this effect earlier in the case.²⁶⁶ The District Court regarded its holding as a narrow one: the Java API elements at issue in *Oracle* were unprotectable methods or systems under § 102(b).²⁶⁷ The law is unsettled as to whether all interfaces, APIs, and command structures should be treated the same.²⁶⁸ The terms "interface," "APIs," and "interoperability" unfortunately do not have fixed and unalterable meanings.²⁶⁹

The District Court regarded Google's interest in enabling technical compatibility of Java programs with the Android platform as a legitimate explanation for Google's use of the Java API packages. It observed, for instance, that "millions of lines of code had been written in Java before Android arrived," and all of them had "necessarily used the `java.package.Class.method()` command format."²⁷⁰ These programs had been developed and were owned by firms other than Google.

In order for at least some of this code to run on Android, Google was required to provide the same `java.package.Class.method()` command system using the same names with the same 'taxonomy' and with the same functional specifications.²⁷¹

Hence, it was not just Google that was interested in enabling Java programs to run on Android; the developers of existing programs would want them to run (or be easily adapted to run) on Android too. The District Court found that Google had used "what was necessary to achieve a degree of interoperability—but no more, taking care, as said before, to provide its own implementations."²⁷² In this respect, the District Court thought *Oracle*

266. *Oracle Am. Inc. v. Google Inc.*, 810 F. Supp. 2d 1002 (N.D. Cal. 2011).

267. *Oracle Am., Inc. v. Google Inc.*, 872 F. Supp. 2d 974, 1000–01 (N.D. Cal. 2012).

268. *Bateman*, 79 F.3d at 1547 (rejecting argument that interface specifications are uncopyrightable as a matter of law). A document that sets forth the component elements of an API with original commentary to aid users would be copyrightable, but the copyright in that document should not be infringed insofar as reuse of the specified API is necessary to achieve program interoperability.

269. *See, e.g.*, BAND & KATOH, INTERFACES 2.0, *supra* note 10, at 41; *see generally* ASHWIN VAN ROOIJEN, THE SOFTWARE INTERFACE BETWEEN COPYRIGHT AND COMPETITION LAW: A LEGAL ANALYSIS OF INTEROPERABILITY IN COMPUTER PROGRAMS (2010).

270. *Oracle*, 872 F. Supp. 2d at 1000.

271. *Id.*

272. *Id.*

was analogous to the *Sega* and *Connectix* cases.²⁷³ *Connectix* seemed especially relevant because that defendant, like Google, had achieved only partial interoperability.²⁷⁴

The CAFC directly challenged Google’s interoperability claim, saying, “Google designed Android so that it would *not* be compatible with the Java platform.”²⁷⁵ There was, in its view, “no evidence in the record that any [fully Java-compatible] app [running on Android] exists and [the trial court] points to no Java apps that either pre-dated or post-dated Android that could run on the Android platform.”²⁷⁶ The CAFC should, however, have either deferred to the District Court’s factual finding that millions of lines of Java code could run on Android, even if that court did not identify specific programs, or have sent the case back to the District Court for further proceedings to develop a fuller record about compatibility issues.²⁷⁷

The CAFC asserted that Google adopted the 37 API packages for Android “to capitalize on the fact that software developers were already trained and experienced in using the Java API packages at issue.”²⁷⁸ That is, Google’s use of Java API method headers would make it easier for Google to attract Java developers to create apps for the Android platform. The CAFC seemed to perceive this as unfair free-riding because it would “accelerat[e]” Google’s development process “by ‘leverag[ing] Java for its existing base of developers.’”²⁷⁹ The CAFC may also have been swayed by the existence of a licensing program that Oracle (and its predecessor Sun) had established to ensure that Java compatibility goals would be maintained, which Google arguably bypassed by using Java packages in Android without a license.²⁸⁰ Oracle claimed that Google’s success with

273. *Id.*

274. *Id.* at 1000–01 (noting that *Connectix* implemented only 137 of 242 Sony BIOS functions).

275. *Oracle Am., Inc. v. Google Inc.*, 750 F.3d 1339, 1371 (Fed. Cir. 2014).

276. *Id.*

277. *See supra* note 240 (concerning the deference issue). The CAFC was incorrect in asserting that no pre-existing Java programs can run on the Android platform.

278. *Oracle*, 750 F.3d at 1371. The CAFC also rejected Google’s argument that the Java API command structure had become unprotectable because it was an industry standard. *Id.* at 1372. Google had made this claim below, but failed to present evidence to support it. *Oracle*, 872 F. Supp. 2d at 999 n.9.

279. *Oracle*, 750 F.3d at 1371.

280. *See id.* at 1350. Google had negotiated with Sun for a time to license the Java technology, but thought it did not need a license to use only parts of the API and not the suite of Java technologies that Sun licensed to some developers. The failure to license the API did not cut against *Accolade* in *Sega*. *See* 977 F.2d 1510, 1514 (9th Cir. 1992).

Android had “fragment[ed]” Java, defeating its purpose, and thwarting Oracle’s ability to develop its own mobile platform.²⁸¹

Let us accept that Google’s motivation in adopting those 37 Java API packages was, in no small part, to facilitate a kind of human interoperability. Oracle has estimated that there are roughly nine million Java programmers in the world.²⁸² All have made significant investments in learning the rules and syntax of Java, as well as in using the method headers and classes of Java API packages, and implementing the Java API in programs that run on a wide variety of machines to accomplish a wide range of tasks. They are familiar with Java commands and command structures. When they write code, they express themselves with Java commands. If Judge Boudin was right to consider user investments in learning the Lotus command structure and constructing macros in the Lotus language as supporting Borland’s right to reimplement the Lotus command hierarchy in QP,²⁸³ other courts should be receptive to user investments in learning to use Java as a consideration that weighs in favor of limiting the scope of copyright in cases such as *Oracle*.

It is understandable that Java programmers want to reuse command names to identify specific functions that their programs are designed to perform. Naming commands is difficult because

[e]ach name must succinctly and intuitively capture the role and function of the concept being named. Because programs use more names than can be reasonably remembered, the names must be

281. These points were heavily emphasized in Oracle’s brief to the CAFC. *See* Opening Brief and Addendum of Plaintiff-Appellant [hereinafter Opening Brief] at 13–29, 51–55, *Oracle*, 750 F.3d 1339 (No. 13-1021), 2013 WL 518611. Sun had, in fact, developed Java Micro Edition platforms for mobile devices that were not fully compatible with the Java language specification. Java MEs have been used on a large number of mobile devices. These platforms did not achieve as much success as Android, in part because of their use of “draconian” security measures and putting interests of telco carriers above interests of users. *See* Matthew Powell, *Why did Nokia fail to compete with Samsung, Apple etc., despite being the giant of the mobile phone industry?*, QUORA (Feb. 15, 2015), <http://www.quora.com/Why-did-Nokia-fail-to-compete-with-Samsung-Apple-etc-despite-being-the-giant-of-the-mobile-phone-industry> [<https://perma.cc/493C-BHPV>].

282. Nikita Salnikov-Tarnovski, *How Many Java Developers Are There in the World?*, DZONE (July 20, 2012), <https://dzone.com/articles/how-many-java-developers-are> [<https://perma.cc/GVL8-GVRV>].

283. *See supra* text accompanying notes 205–208; *see also* John Bergmayer, *Compatibility is About Competition, Too*, PUBLIC KNOWLEDGE (Feb. 26, 2015), <https://www.publicknowledge.org/news-blog/compatibility-is-about-competition-too> [<https://perma.cc/Z9FY-J943>] (“It’s about allowing developers to make the most of their skills and *their* code, not about Android trying to take something away from Oracle’s version of Java.”).

systematized to be easy to reconstruct and interpret later. This, in turn, will affect how easily you or others can understand, fix, and extend that same code months or years later. As a result, bad names create buggy code, and good names can deliver tremendous boosts to team productivity.²⁸⁴

The Java API systematized names in a very efficient way, which is why the language and the API has been so widely used. The creators of Java can, of course, take some credit for this success, but they cannot take all of it, just as they cannot claim ownership over consumer choice and investment in determining the popularity of any given product.²⁸⁵

Oracle contended that Google could have used different names for method headers for specific functions in support of its claim that Google did not have to copy the method headers that Sun’s engineers developed.²⁸⁶ In effect, Oracle was arguing that Google should have forced Java programmers to learn a new dialect of Java to write apps for the Android platform. Yet, had Google developed a new Java dialect, Google would have confused Java programmers who would have had to relearn to program in Java in a different way for Android than for other devices. This would have caused a much more serious fragmentation of Java than Google caused by adopting only 37 of the Java packages instead of all of them. So if Oracle really valued the integrity of Java, it would not have wanted Google to develop alternative method headers in the Java language.

The best way to have avoided the fragmentation of Java, of which Oracle complained, would have been for Google to adopt all 166 of the API packages instead of just 37 of them. Had Google adopted all 166 packages, a more complete interoperability of Java programs on Android would presumably have been achieved. It would also have made the *Oracle* case more like *Sega*.

The Java API may be a much more complex procedure than the Genesis interface at issue in *Sega*, but the complexity of a procedure does not change its essential character. Under Ninth Circuit precedents,²⁸⁷ the Java API and its constituent parts would be unprotectable under § 102(b) insofar as its use

284. Excerpt from an online discussion in answer to the question “Why is naming things hard in computer science and how can it be made easier?”, QUORA (answer updated Feb. 17, 2014), <http://quora.com/Why-is-naming-things-hard-in-computer-science-and-how-can-it-be-made-easier> [<https://perma.cc/AW92-BEEZ>].

285. See generally Michal Shur-Ofry, *Popularity as a Factor in Copyright Law*, 59 U. TORONTO L.J. 525 (2009).

286. See Opening Brief, *supra* note 281, at 32–33, 51–52.

287. See *Sony Comput. Entm’t, Inc. v. Connectix Corp.*, 203 F.3d 596, 602–03 (9th Cir. 2000); *Sega Enters. Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1526 (9th Cir. 1992).

was needed to achieve interoperability, as the District Court in *Oracle* held.²⁸⁸

Even if the Java API as a whole might be an unprotectable system, method, or procedure under Ninth Circuit precedents, it is unclear whether or why the structure of 37 of the Java API packages should be regarded as constituting a system or method of operation. The District Court did not explain why a subset of the API should be regarded as a system or method as well.

Yet, if we accept that the Java API as a whole was an unprotectable system or procedure, it would be logical to conclude that its constituent parts should be unprotectable as well. The Ninth Circuit did not treat *Connectix*'s partial implementation of the Sony PlayStation interface as any less of a procedure than in *Sega*.²⁸⁹ Consider also that a second comer would be free to copy 6000 items from an uncopyrightable compilation of 60,000 white pages listings of a telephone directory. Baker did not copy Selden's forms exactly, but rather adapted the forms so that it would be easier for accountants to enter data on an ongoing basis.²⁹⁰ Baker thus used some of the Selden system in his forms, but he did not become an infringer for not using all of it. Consider further that Android is a special purpose platform for mobile devices, a type of platform that was not in contemplation when Java was developed. The success of Java has mainly been achieved with server-side systems and large-scale enterprise software that can run on multiple machines. It stands to reason that a special purpose platform for mobile devices would need to have some of the same, but also some different, functionalities that its API would need to accommodate. This could explain why Google used only some, but not all, of the Java API

288. *Oracle*, 872 F. Supp. 2d at 997–1000; see *Atari Games Corp. v. Nintendo of Am., Inc.*, 897 F.2d 1572, 1575 (Fed. Cir. 1990) (“When the questions on appeal involve law and precedent on subjects not exclusively assigned to the Federal Circuit, the court applies the law which would be applied by the regional circuit.”) (quoted in *Oracle*, 750 F.3d at 1353). Yet, the CAFC cited to its opinions in *Atari Games* ten times as the most relevant “Ninth Circuit” precedents. It cited the *Apple v. Microsoft* decision only once for a minor point, even though that decision is the Ninth Circuit’s principal software copyright decision. The Ninth Circuit has cited the *Atari Games* decision only twice in passing in the last decade in the Westlaw database, whereas it has cited *Apple v. Microsoft* eighty-one times.

289. See *Connectix*, 203 F.3d at 599, 609 (noting lack of full compatibility), 602–03, 607 (Sony interface held to be § 102(b) procedure); see also Google Appellee Brief, *supra* note 212, at 37 (noting that *Connectix* had used 137 of 242 elements of the Sony PlayStation interface).

290. Samuelson, *Baker v. Selden*, *supra* note 155, at 164.

packages.²⁹¹ Even if avoiding fragmentation of Java was a commercial and competitive objective for Oracle, that consideration has no bearing on whether the Java API packages were unprotectable under § 102(b).

Until the *Oracle* decision, the law was settled that program interfaces necessary to achieving interoperability among programs or with hardware were unprotectable by copyright law because of their inherent functionality. The *Oracle* decision incorrectly interpreted Ninth Circuit precedents, including the significance of § 102(b) in cases involving program APIs. Healthy competition and innovation in the software industry has depended for decades on the ability of second comers to build new programs that interoperate with older ones, even when the initial developer of the API at issue has not consented to this use.

IV. FUNCTIONALITY AND EXPRESSION SOMETIMES MERGE IN SOFTWARE CASES

Copyright law has long recognized that when there is only one or a small number of ways to express an idea or function, copyright protection will be withheld to any expression that has merged with the idea or function.²⁹² Software APIs necessary for achieving interoperability are among the computer program innovations in which function and expression effectively merge. APIs specify a system of rules and procedures to which other programs must strictly conform in order to attain compatibility with an existing program. The fact that software engineers might *ex ante* have had numerous choices in how to design an API does not make the API copyrightable because that API significantly constrains design choices of subsequent programmers.

Section A discusses the origins and scope of copyright's merger doctrine. Section B considers the role of the merger doctrine in architectural work and software cases. Section C reviews software copyright cases in which a first programmer's design choices constrained the design choices of subsequent programmers, particularly when reuse of an API is necessary to achieving interoperability. Section D recognizes that on some occasions, function and expression may merge as to program code. Section E identifies several errors in the CAFC's interpretation of the merger doctrine in its *Oracle* decision.

291. It is common for technologists to tinker with existing technologies and adapt them to different uses than their manufacturers intended. Considerable innovation has occurred from these adaptations. *See generally* ERIC VON HIPPEL, *DEMOCRATIZING INNOVATION* (2005).

292. *See infra* text accompanying notes 293–301; *see also* Pamela Samuelson, *Reconceptualizing Copyright's Merger Doctrine*, 63 J. COP. SOC'Y U.S.A. 417, 417 (2016).

A. ORIGINS OF THE MERGER DOCTRINE

The challenge of distinguishing expression and functionality in copyrighted works has been with U.S. copyright law for a very long time. The Supreme Court first addressed the issue in *Baker* when holding that Selden's copyright extended to his *explanation* of his bookkeeping system, not to the bookkeeping *system* itself.²⁹³ The Court characterized Selden's forms as "necessary incidents" to this useful art, and since practicing that system required use of Selden's forms, the forms were uncopyrightable as well.²⁹⁴

Courts and commentators often point to the "necessary incidents" language when speaking of *Baker* as having originated copyright's merger doctrine.²⁹⁵ The selection and arrangement of columns and headings in Selden's forms had, in this conception of *Baker*, in effect, merged with the system and the method of operation that Selden anticipated accountants would carry out when using the forms.²⁹⁶

293. 101 U.S. 99, 101 (1879). *Baker* is best understood as holding that systems, methods of operation, and other useful arts are excluded from the scope of copyright protection, a doctrine that is now codified in § 102(b). See Samuelson, *Why Copyright Excludes Systems*, *supra* note 124, at 1931–42 (explaining *Baker* and its progeny on this point). Contrary to common perception, *Baker* is not the source, or even as a classic statement, of the idea/expression distinction, although many courts characterize in this way. See *id.* at 1924–36 (explaining that the exclusion of ideas from the scope of copyright predated *Baker* and was not at issue in that case).

294. *Baker*, 101 U.S. at 103. The Nimmer treatise has asserted that Baker's forms did not infringe because they differed in some respects from the Selden forms. NIMMER ON COPYRIGHT, *supra* note 10, at § 2.18[B][1]. The relevant forms can be found in Samuelson, *Baker v. Selden*, *supra* note 155, at 170–71. The logic of the Court's decision, however, supports the view that exact copying of the forms would not infringe. See, e.g., BENJAMIN KAPLAN, AN UNHURRIED VIEW OF COPYRIGHT 63–64 (1966); HORACE G. BALL, THE LAW OF COPYRIGHT AND LITERARY PROPERTY 111–12, 125–28 (1944) (discussing *Baker* and its progeny as precedents for the unprotectability of systems of business, plans of instruction, or methods of practicing an art or playing a game); ARTHUR W. WEIL, AMERICAN COPYRIGHT LAW 193–94 (1917) (citing *Baker* and its progeny as precluding copyright protection for plans, methods, and arts).

295. Among the many cases that cite to *Baker* as the origin of the merger doctrine is *Arica Inst. v. Palmer*, 970 F.2d 1067, 1076 (2d Cir. 1992). The Nimmer treatise is among the many sources that so credit *Baker*. NIMMER ON COPYRIGHT, *supra* note 10, at §2.18. I have elsewhere argued that *Baker* did not, in fact, give birth to the merger doctrine, but rather to the exclusion of procedures, processes, systems, and methods of operation now codified in § 102(b). See Samuelson, *Reconceptualizing Merger*, *supra* note 292, at 419–22. In numerous cases, merger and § 102(b) defenses have been treated interchangeably. See *id.* at 451–53.

296. See, e.g., Burk, *supra* note 206, at 591 (Selden's "accounting forms were the only way to express the accounting system [so] that the idea and expression had merged.").

After *Baker*, courts in the U.S. have mainly struggled with the problem of the separability or merger of functionality and expression when deciding whether pictorial, graphic, or sculptural (PGS) works are protectable by copyright law.²⁹⁷ The 1976 Act allows PGS works to be copyrighted insofar as they embody original expression that “merely [] portray[s] the appearance of [an] article or [] convey[s] information.”²⁹⁸ If such works have “an intrinsic utilitarian function” that goes beyond portraying an appearance or conveying information, they are disqualified from U.S. copyright protection.²⁹⁹ If the design of a PGS work “incorporates pictorial, graphic, or sculptural features that can be identified separately from and are capable of existing independently of, the utilitarian aspects,” then the work is copyrightable.³⁰⁰ If functionality and expression are inseparable (i.e., merged), then no matter how creative its design, the PGS work is unprotectable by copyright law.³⁰¹

Mazer v. Stein is a classic copyright case in which expression and functionality were separable.³⁰² Stein’s statuette of a Balinese dancer was eligible for copyright protection because the statuette was a work of art that existed independently from the lamp.³⁰³ It did not matter that the statuette was being commercially exploited as the base for Stein’s lamps because the lamp did not function any better or worse for having the statuette as its base. Jewelry, fabric designs, and dolls are among the other types of intellectual creations that, having satisfied the separability criterion, enjoy copyright protection.³⁰⁴

The overall design of a chair or automobile, by contrast, may well have an aesthetic character, but any expressive aspects of these creations cannot be separated from (that is, they are merged with) their utilitarian functions.³⁰⁵ The merger of functionality and expression in chairs,

297. See, e.g., *Kieselstein-Cord v. Accessories by Pearl, Inc.*, 632 F.2d 989 (2d Cir. 1980) (separable expression in belt jewelry); *Carol Barnhart Inc. v. Economy Cover Corp.*, 773 F.2d 411 (2d Cir. 1985) (no separable expression in mannequin).

298. 17 U.S.C. § 101 (definition of “useful article”).

299. *Id.* (definition of “useful article”).

300. *Id.* (definition of PGS works).

301. See, e.g., *Burk*, *supra* note 206, at 591 (characterizing the useful article/PGS rule as a kind of merger doctrine).

302. 347 U.S. 201 (1954).

303. See *id.* at 215–18.

304. See, e.g., *Kieselstein-Cord v. Accessories by Pearl, Inc.*, 632 F.2d 989, 993 (2d Cir. 1980) (finding that the conceptually separable artistic elements of belt buckles are copyrightable).

305. See H.R. REP. NO. 94-1476, at 55 (1976) (giving numerous examples of designs of useful articles which, even if original, are unprotectable under U.S. copyright law).

automobiles, and other useful articles results in no copyright protection whatsoever in the U.S.³⁰⁶ Moreover, to ensure that copyrights in technical drawings or depictions of functional creations do not indirectly undermine the no-copyright-for-functionality rule, the 1976 Act and case law clarify that any copyright in the drawing does not extend protection to the technical content or functional creations depicted therein.³⁰⁷

B. THE ROLE OF THE MERGER DOCTRINE IN ARCHITECTURAL WORK AND SOFTWARE CASES

Congress has created two limited exceptions to the no-copyright-for-functional-creations rule: one for computer programs and another for architectural works.³⁰⁸ By designating programs and architecture as copyrightable subject matters, Congress did not jettison the longstanding rule that copyright does not protect functionality. Instead, it shifted the problem of assessing whether expression and functionality are separable or merged so that the question is, generally speaking, no longer about eligibility for copyright protection,³⁰⁹ but rather about assessing the proper scope of protection for such works.

The courts have explicitly recognized the relevance of the merger doctrine as a limit on the scope of copyright protection for functional design elements of architectural works. In *Zalewski v. Cicero Builder Dev., Inc.*, for example, the Second Circuit recognized that efficiency considerations may significantly constrain the design of buildings, as well as computer programs, and adapted the filtration factors first articulated in *Altai* to apply

306. See, e.g., *Brandir Int'l, Inc. v. Cascade Pacific Lumber Co.*, 834 F.2d 1142, 1147 (2d Cir. 1987) (ribbon design for bicycle rack held unprotectable by copyright law as useful article lacking separable expression).

307. 17 U.S.C. § 113(b); see, e.g., *Niemi v. American Axle Mfg.*, No. 05-74210, 2006 WL 2077590 (E.D. Mich. July 24, 2006) (unauthorized manufacture of machine did not infringe copyright in drawing); *Eliya, Inc. v. Kohl's Dept. Stores*, 82 U.S.P.Q.2d 1088 (S.D.N.Y. 2006) (manufacture of shoe did not infringe copyright in drawing); *Fulmer v. United States*, 103 F. Supp. 1021, 1022 (Ct. Cl. 1952) (copyright in drawing of parachute design did not give its author an exclusive right to make parachutes like that depicted in the drawing).

308. See *supra* note 2 (discussing the Congressional decision to extend copyright protection to programs). Congress added architectural works to copyright subject matter in 1990. Copyright Improvements Act of 1990, Pub. L. No. 101-650, Tit. VII, 104 Stat. 5133 (codified in various sections of 17 U.S.C.). Copyright Office regulations state that structures, such as “bridges, cloverleaves, dams, walkways, tents, recreational vehicles, mobile homes and boats,” are ineligible for copyright protection. 37 C.F.R. § 202.11(d). Functionality predominates over aesthetics in the design of these structures.

309. On rare occasions, merger of function and expression may preclude copyright protection for some program code. See *infra* notes 318–319 and accompanying text.

in architectural work cases.³¹⁰ To the extent that “design elements [are] attributable to building codes, topography, structures that already exist on the construction site, or engineering necessity[, they] should therefore get no protection.”³¹¹ The court observed that methods of construction and good engineering practices were likewise unprotectable by copyright.³¹² It further noted that “functional aspects of a work are governed by patent law, not copyright law.”³¹³ Zalewski’s infringement claim failed because “[t]here are only so many ways to arrange four bedrooms upstairs and a kitchen, dining room, living room, and study downstairs” for colonial style homes.³¹⁴

Copyright protection in an architectural work extends to the aesthetic design of a building’s exterior and interior, but not to utilitarian features such as the layout of electrical, heating, or plumbing infrastructures.³¹⁵ There may, *ex ante*, be many different ways to design the wiring, heating or plumbing systems in a building, but no court would be confused into thinking that a particular layout was protectable expression in the copyrighted building. Those systems lie outside the scope of copyright in any drawing of the building or in the building itself, no matter how much creative effort went into the designs for these systems and no matter how many design choices the architect might have had at the outset. These are functional elements of the architectural design because any creativity in the layout of wires, heating, or plumbing systems is too interconnected with functionality to be part of the expression of copyrighted building.³¹⁶

It would be a doctrinal advance in copyright law if courts articulated a function/expression merger doctrine, as such, in computer program cases.³¹⁷

310. *Zalewski v. Cicero Builder Dev., Inc.*, 754 F.3d 95, 105 (2d Cir. 2014) (citing and quoting from *Altai*).

311. *Id.*

312. *Id.* at 105–06.

313. *Id.* at 106 n.19.

314. *Id.* at 107. Note that the court did not invalidate Zalewski’s copyright, but regarded the scope of that copyright as sufficiently thin that Cicero did not infringe. *Id.* at 106.

315. *See, e.g.*, COPYRIGHT CLAIMS IN ARCHITECTURAL WORKS, COPYRIGHT OFFICE CIRCULAR 41, at 1–2 (The scope of copyright in copyrighted buildings does not extend to standard configurations of spaces and individual elements such as doors and windows, or “functional elements whose design or placement is dictated by utilitarian concerns.”).

316. The existence of alternative ways to achieve the same functionality is routinely disclosed in utility patents which differentiate the claimed invention from the prior art. *See infra* notes 335–37 and accompanying text.

317. Some courts speak of a process/expression distinction in computer program cases. *See, e.g.*, *Gates Rubber Co. v. Bando Chem. Indus.*, 9 F.3d 823, 836–37 (10th Cir. 1993). This implicitly lays the groundwork for recognition of a process/expression merger doctrine.

The *Altai* decision came close to doing this when the Second Circuit asserted that the merger doctrine precluded copyright protection for efficient design elements of programs.³¹⁸ A computer program function/expression merger doctrine could supplement the function/expression merger rule that disqualifies many aesthetic designs of articles of manufacture from copyright protection.³¹⁹ It would also complement the fact/expression merger doctrine recognized in recent cases.³²⁰ Merger of ideas and expressions is, as it happens, just one type of merger that courts should (and do) recognize.³²¹

Given that programs are utilitarian works that courts often say should enjoy only a thin scope of copyright protection,³²² it is curious that courts have rarely articulated or applied a function/expression merger doctrine, as such, in software infringement cases. Courts' reluctance to do this may stem from the misleading nature of the literary work metaphor that so often pervades copyright discourse about computer programs.³²³ This metaphor obscures the deeply functional nature of programs, which have a higher quantum of utilitarian elements as compared with conventional literary works such as novels and plays. While it may sometimes be difficult to distinguish the ideas and the expressions in conventional literary works, there is generally no need to try to separate literary expressions in these works from functionalities because they generally have none.³²⁴

See also Mitek Holdings, Inc. v. Arce Eng'g Co., 89 F.3d 1568, 1556 n.19 (11th Cir. 1996); Atari Games Corp. v. Nintendo of Am., Inc., 975 F.2d 832, 839–40 (Fed. Cir. 1992) (recognizing the possibility of process/expression merger).

318. Comput. Assocs. Int'l, Inc. v. Altai, Inc., 982 F.2d 693, 707–09 (2d Cir. 1992).

319. *See supra* text accompanying notes 298–307. While usually called the separability doctrine, it effectively is a merger doctrine because the opposite of separable is merged.

320. *See, e.g.*, N.Y. Mercantile Exch., Inc. v. Intercontinental Exchange, Inc., 497 F.3d 109, 116–19 (2d Cir. 2007) (NYMEX's creation of settlement prices for futures contracts treated as fact/expression merger); *Banxcorp. v. Costco Wholesale Corp.*, 978 F. Supp. 2d 280, 308 (S.D.N.Y. 2013) (merger doctrine precludes copyright protection of interest rate averages, even though some variation in expression of this fact is possible); *see also* Veeck v. Southern Building Code Congress Int'l, Inc., 293 F.3d 791, 800–02 (5th Cir. 2002) (holding that enacted building codes were unprotectable facts under the merger doctrine). Shubha Ghosh has argued for treating enacted codes as instances of function/expression merger. *See* Shubha Ghosh, *Legal Code and the Need for a Broader Functionality Doctrine in Copyright*, 50 J. COP. SOC'Y 71, 104–08 (2003).

321. *See* Samuelson, *Reconceptualizing Merger*, *supra* note 292, at 438–42.

322. *Apple Computer Inc. v. Microsoft Corp.*, 35 F.3d 1435, 1444 (9th Cir. 1994); *Altai*, 982 F.2d at 712; *Sega Enters. Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1524 (9th Cir. 1992).

323. *See, e.g.*, Weinreb, *supra* note 10, at 1167–70.

324. An example of a literary work with separable functionalities is *Red Sparrow* (2013) by Jason Matthews, a spy novel with recipes at the end of each chapter; the book's

Computer programs, by contrast, are virtual machines, “machines . . . that have been constructed in the medium of text.”³²⁵ This characterization of programs is not just a metaphor; this is what programs actually are. Characterizing them as “literary works” is not wrong, given the capacious statutory definition of that term,³²⁶ but it is seriously incomplete because it obscures the deeply functional nature of programs and program designs, and ignores the functional behavior that motivates people to purchase them.

Programs are products of an industrial design process.³²⁷ “[C]reating a program is a process of building and assembling functional elements.”³²⁸ Unlike physical machines that are constructed from gears, wires, screws, and other hardware components, “programs are built from information structures, such as algorithms and data structures.”³²⁹ They are industrial compilations of applied know-how.³³⁰

Neither copyright nor patent law has conventionally extended protection to these kinds of innovations,³³¹ although industrial compilations of applied know-how may be and often are protected as trade secrets. Indeed, internal design elements of programs are most often and best protected as secrets, making both copyright and patent protection less needed for nonliteral elements of software. While industrial design elements of programs can and sometimes are reverse-engineered, reverse-engineering of program code does not generally enable the engineer to engage in market-destructive appropriations of program know-how because this method of discovery of program internals is difficult, expensive, and time-consuming.³³²

Command structures and APIs, along with program behavior, are industrial compilations of applied know-how that often cannot be kept secret.³³³ Command structures may be visible when users run the program, as in *Borland*, or they may be published, as were the API packages in *Oracle*. But that does not change their essential nature. Functionality and expression are so tightly coupled (i.e., merged) in command structures and APIs that they should lie outside the scope of copyright protection in

copyright would not extend to the recipes. *See* *Publications Int’l Ltd. v. Meredith Corp.*, 88 F.3d 473, 480–81 (7th Cir. 1996) (finding recipes in cookbook not entitled to copyright protection).

325. Samuelson et al., *Manifesto*, *supra* note 10, at 2316; *see also id.* at 2316–32 (discussing the nature of programs and their designs).

326. 17 U.S.C. § 101 (definition of “literary works”).

327. Samuelson et al., *Manifesto*, *supra* note 10, at 2329–30.

328. *Id.* at 2321.

329. *Id.*

330. *Id.* at 2326–32.

331. *Id.* at 2342–56.

332. *Id.* at 2389–93.

333. *Id.* at 2356–64.

programs. After all, command structures and APIs are carefully engineered to identify the functions of which programs are capable and the manner in which information must be configured to be exchanged across program boundaries so that programs are able to function properly. As the *Borland* and *Sega* decisions, among others, demonstrate, programmers can generally develop independent implementations of methods, procedures, and processes, such as APIs and command structures, without infringing copyrights. The CAFC failed to recognize this in *Oracle*.

Oracle is among the many plaintiffs in software copyright cases that have characterized computer programs as literary works and insisted that merger is relevant only when the plaintiff had no design alternatives, as the Third Circuit said in *Whelan* and the CAFC in *Oracle*.³³⁴ This approach is unduly restrictive of programmer reuses of functional designs that could contribute to ongoing competition and follow-on innovation in the software industry. Other courts have recognized that there may be more than one method or system that can accomplish a program task, and patents often make reference to prior art that performed the same function in a different way.³³⁵ To be consistent with Supreme Court decisions in other IP cases, the existence of alternative ways to perform a function should not be the sole criterion for whether to treat that type of nonliteral element of a program as expressive, as it was in the CAFC's *Oracle* decision.³³⁶ Also relevant are purpose, cost, and use characteristics of the creation.³³⁷ This is especially true for interfaces necessary for interoperability.

334. See *supra* text accompanying notes 28, 65.

335. See, e.g., *Bateman v. Mnemonics, Inc.*, 79 F.3d 1532, 1546, n.29 (11th Cir. 1996) (“The availability of alternatives should not be determinative in distinguishing elements of a computer program that are expression and those that are unprotectable under 102(b). Generally, there is more than one method of operation or process that can be used to perform a particular computer program function . . . Patents routinely recite prior methods or systems of performing the same function in distinguishing the claimed invention from the prior art.”); see also *Bikram’s Yoga Coll. of India, L.P. v. Evolution Yoga, LLC*, 803 F.3d 1032, 1042 (9th Cir. 2015) (“[T]he possibility of attaining a particular end through multiple different methods does not render the uncopyrightable [yoga sequences] a proper subject of copyright.”).

336. The existence of design alternatives may similarly be a factor in deciding whether a design is too functional to be protected as trade dress, but the Supreme Court has rejected this as a sole criterion for judging the nonfunctionality of trade dress in *TrafFix Devices, Inc. v. Marketing Display Inc.* See 532 U.S. 23, 27–32 (2001). The implications of *TrafFix* for software copyright cases are discussed *infra* text accompanying notes 424–427. See also Weinreb, *supra* note 10, at 1170 (“[I]f this rubric [of other possible design choices] is used, copyright effectively absorbs the whole of patent.”).

337. See, e.g., *TrafFix*, 532 U.S. at 31–33; *Kellogg Co. v. Nat’l Biscuit Co.*, 305 U.S. 111, 122 (1938) (rejecting claim of trademark in pillow shape of Shredded Wheat biscuits,

C. MERGER MAY BE FOUND WHEN A PLAINTIFF’S DESIGN CHOICES
SERVE AS CONSTRAINTS ON THE CHOICES AVAILABLE TO SECOND
COMERS

Baker teaches that expression and functionality are sometimes too closely intermixed in some textual works for copyright protection to be available to those elements. When Selden first devised the bookkeeping system embodied in his book, there were other ways to keep books for various accounts, but he was dissatisfied with them. In designing forms for his new system, he wasn’t completely free to arrange columns and headings as he wished because state law required all bookkeeping forms to have columns captioned “date,” “no.,” “to,” “for,” and “by.”³³⁸ Selden’s forms complied with this mandate in the five farthest left-hand columns. But he was not constrained in other design decisions for arranging columns and headings in the forms.

Selden believed the selection and arrangement of columns and headings in the new forms was highly creative, as the preface to his book revealed: “To greatly simplify the accounts of extensive establishments doing credit business, and embracing an almost infinite variety of transactions would be a masterly achievement, worthy to be classed among the greatest benefactions of the age.”³³⁹ So even though Selden could easily meet a creativity-based originality standard, his forms were nonetheless unprotectable by copyright law because the selection and arrangement of the columns and headings were “necessary incidents” to the system he devised.³⁴⁰

This aspect of *Baker* demonstrates that the merger doctrine may apply even when the copyright claimant had other choices when first developing a creative procedure, process, system, or method of operation. Selden’s copyright was, of course, valid, but it provided protection only to his explanation of the system, not to the system and the forms that instantiated it. Unlike most useful arts, which are embodied in metal or other materials, Selden’s useful art was embodied in a book (or in today’s parlance, a literary work). But the Court said in *Baker*: “[T]he principle is the same in all. The description of the [useful] art in a book, though entitled to the benefit of

despite existence of design alternatives, because “[t]he evidence is persuasive that this form is functional—that the cost of the biscuit would be increased and its high quality lessened if some other form was substituted for the pillow-shape”).

338. Samuelson, *Baker v. Selden*, *supra* note 155, at 168.

339. *Id.* at 160 (quoting from the Supreme Court record).

340. *Baker v. Selden*, 101 U.S. 99, 103 (1879).

copyright, lays no foundation for an exclusive claim to the art itself.”³⁴¹ The creativity required to develop the system and to devise forms to implement it did not make the system or the forms copyrightable because of the merger of function and expression. By not protecting Selden’s system and the forms embodying it, there was breathing room for later bookkeepers to continue to evolve the useful art of bookkeeping and devise new forms that further improved upon Selden’s innovation, as Baker himself did.³⁴²

Courts have followed *Baker*’s admonitions in many subsequent non-software cases.³⁴³ The pertinence of *Baker* for assessing the scope of copyright in software was importantly acknowledged in *Altai*. *Altai* directly invoked *Baker* and the merger doctrine in explaining why efficient functional design elements of programs lie outside of the scope of copyright protection available to programmers.³⁴⁴ The Second Circuit observed that “[i]n the context of computer program design, the concept of efficiency is akin to deriving the most concise logical proof or formulating the most succinct mathematical computation.”³⁴⁵ The more efficient a nonliteral element of a program is, the closer it approximates a merger of idea and expression. The court recognized that “hypothetically, there might be a myriad of ways in which the programmer may effectuate certain functions within a program . . . [but] efficiency considerations may so narrow the practical range of choice as to make only one or two forms of expression workable options.”³⁴⁶

The court in *Altai* was implicitly concerned that the first software developer to devise an efficient functional design for a program should not be able to get 95 years of protection for it and force all other programmers to adopt less efficient solutions. It recognized that “[e]fficiency is an industry-wide goal” for software developers.³⁴⁷ *Altai* directed that efficient design elements of programs be filtered out of consideration before doing the final step in infringement analysis, even though other design choices might have been available.

Elements of programs that are “dictated by external factors” must also be filtered out.³⁴⁸ The Second Circuit in *Altai* identified several types of

341. *Id.* at 105 (emphasis added).

342. *See* Samuelson, *Baker v. Selden*, *supra* note 155, at 193; *see also id.* at 169 n.76.

343. *See* Samuelson, *Why Copyright Excludes Systems*, *supra* note 124, at 1936–44 (discussing the cases that followed *Baker*).

344. *Comput. Assocs. Int’l, Inc. v. Altai, Inc.*, 982 F.2d 693, 707–08 (2d Cir. 1992).

345. *Id.* at 708.

346. *Id.*

347. *Id.*

348. *Id.* at 710.

constraints that might limit the design choices of programmers, including “compatibility requirements of other programs with which a program is designed to operate in conjunction.”³⁴⁹ Although the court referred to the scenes a faire doctrine as a justification for treating compatibility as an external constraint on programmer design decisions,³⁵⁰ to the extent it used “dictated by” language,³⁵¹ the more pertinent doctrine is merger. *Altai* and its progeny regard compatibility as an external factor constraint on design decisions of defendants who are developing programs to interoperate with existing software and hardware.

The CAFC’s *Oracle* decision interpreted this external factors constraints category too narrowly. That court considered *Altai* to have directed the filtration of elements dictated by external constraints only insofar as the constraints limited the plaintiff’s design choices.³⁵² Having determined that Sun’s engineers were not constrained in designing the Java API packages, the CAFC regarded Google’s external constraints argument to be unpersuasive.³⁵³ However, the Second Circuit did not so limit the external constraints category in *Altai*, and in the twenty-three years since

349. *Id.* at 709–10.

350. *See id.* The scenes a faire doctrine is similar to merger in limiting the scope of copyright protection, but it is more focused on whether elements in a protected work are common in works of that kind, not whether they are “dictated” by functional considerations, as merger is in software cases.

351. *See id.*

352. *Oracle Am., Inc. v. Google Inc.*, 750 F.3d 1339, 1370 (Fed. Cir. 2014). The CAFC cited *Dun & Bradstreet Software Servs., Inc. v. Grace Consulting, Inc.*, 307 F.3d 197, 215 (3d Cir. 2002), in support of this interpretation of *Altai*. But in *Grace*, the Court of Appeals for the Third Circuit considered defendant Grace to be a very bad actor who breached license agreements, copied and modified D&B software, and misappropriated its trade secrets. *See Grace*, 307 F.3d 197. In this context, the Third Circuit’s rejection of the defendant’s claim that its design choices were constrained by what D&B had done should be given little weight.

353. *See Oracle*, 750 F.3d at 1370–71. The CAFC fell back on a *Whelan*-like “is there any other way to do it?” approach to assessing whether the Java API packages were expressive. *Id.* But the CAFC was mistaken on this point. Joshua Bloch, one of the Sun engineers involved in designing the Java API, reports that considerable effort went into developing the Java API to faithfully reimplement the syntax and semantics of Perl so that Perl-trained engineers could more easily work in Java. Sun ran a battery of 30,000 tests to ensure that the Java implementation was consistent with Perl’s. Bloch states that the “Java APIs included many preexisting APIs and have since the earliest days of the platform. Many of the original Java APIs were pretty much copied from C to make it easy for C programmers to make the transition.” Communication with Joshua Bloch, Sept. 28, 2015 (on file with author).

Altai, numerous cases have treated external constraints as affecting the defendant's as well as the plaintiff's design choices.³⁵⁴

Anyone who develops an API for achieving program interoperability is, in effect, creating a constraint on his own subsequent design decisions. At the same time, though, that same API developer is also creating constraints on the design choices of all others who want to develop programs to interoperate with his platform, as the Ninth Circuit implicitly recognized in *Sega*.³⁵⁵ When Sega initially developed the interface for its Genesis platform, it had many choices about how to construct that interface. But once that interface existed, Sega and its licensees had to conform to it when they made games for the Genesis. Accolade similarly could not make its videogames run on the Genesis platform without reimplementing the Sega interface in its program. That interface constituted the “functional requirements for achieving compatibility.”³⁵⁶ Although the Ninth Circuit characterized “interface procedures” of the Sega program as unprotectable under § 102(b), merger would have been a reasonable alternative ground.³⁵⁷ Any arguably expressive elements in the Sega interface would be merged with its functionality because third party software cannot execute on the Sega platform unless the interface components exactly conform to the rules that Sega established when designing the interface.³⁵⁸

D. SOMETIMES PROGRAM FUNCTIONS MERGE WITH PROGRAM CODE

The interface at issue in *Altai* was a nonliteral element of a program that CA alleged *Altai* infringed. Sometimes, however, literal copying of code is necessary to achieve interoperability. The Ninth Circuit recognized this in *Sega* because it excused Accolade from infringement for copying a segment of Sega code that was essential to achieving interoperability.³⁵⁹ The

354. *Comput. Assocs. Int'l, Inc. v. Altai, Inc.*, 982 F.2d 693, 709–10 (2d Cir. 1992) (using “dictated by” twice in discussing filtration of external factors). *Sega* and *Borland* are among the decisions in which the defendant's design choices were constrained by the plaintiff's earlier choices. *See also* Samuelson, *Reconceptualizing Merger*, *supra* note 292, at 442–44.

355. *See Sega Enters. Ltd. v. Accolade, Inc.*, 977 F.2d 1510 (9th Cir. 1992).

356. *See id.* at 1522.

357. *Id.* Note that the “functional requirements for . . . compatibility” phrase uses the language of merger. *See also id.* at 1524 (recognizing that “necessary incidents” components of programs are unprotectable, as are program elements “dictated by the function to be performed”).

358. Connectix faced similar constraints when reimplementing the Sony PlayStation interface for its emulation software. *See Sony Comput. Entm't, Inc. v. Connectix Corp.*, 203 F.3d 596, 602–07 (9th Cir. 2000).

359. *Sega Enters. Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1524, 1530–32 (9th Cir. 1992); *see also Oracle Am., Inc. v. Google Inc.*, 872 F. Supp. 2d 974, 995 (N.D. Cal. 2012)

Eleventh Circuit in *Bateman* acknowledged this when ruling that a lower court erred in instructing a jury that only nonliteral elements of a program interface needed for achieving compatibility could be filtered out in applying the AFC test, thus recognizing that sometimes exact copying is truly essential to interoperability.³⁶⁰ The Sixth Circuit in *Lexmark* likewise struck down a lower court ruling of infringement because exact copying of a short Lexmark program installed in that firm's printer cartridges was a necessary incident to interoperation with the Lexmark printer.³⁶¹ The Sixth Circuit expressly relied on merger as the basis for ruling that Lexmark's printer code was ineligible for copyright protection because it was needed to achieve interoperability.³⁶²

Read together, the *Altai*, *Sega*, *Connectix*, *Bateman*, and *Lexmark* decisions not only recognize, but directly hold, that there is a compatibility exception to copyright protection for computer programs when reuse of interface components are necessary for interoperability.³⁶³ This is true whether the interface is embodied in a program or merely specified in a document that is not itself a program. The distinction between APIs and implementations of APIs in independently written code is fundamental in the field of computing.³⁶⁴ Copyright law protects the code that implements an API, but does not protect the API insofar as it is necessary to enabling

(quoting *Sega*, 977 F.2d at 1524 n.7, as excusing exact copying of a short portion of the Sega code because of its functionality).

360. 79 F.3d 1532, 1546–47 (11th Cir. 1996).

361. *Lexmark Int'l, Inc. v. Static Control Components*, 387 F.3d 522, 542 (6th Cir. 2004) (“[I]f any single byte of the Toner Loading Program is altered, the printer will not function . . .”).

362. *Id.*

363. The court in *Bateman* was unwilling to rule that all APIs were unprotectable as a matter of law, but it directed the lower court to instruct the jury that insofar as program code or APIs were necessary to interoperability, they were not within the scope of protection available to programs. 79 F.3d 1532, 1546–57 (11th Cir. 1996). The interpretation of *Borland* discussed above would also support a compatibility exception. See *supra* Section III.B; see also Burk, *supra* note 206, at 591 (suggesting that *Borland* should have been decided on merger of functionality and expression grounds). In *Oracle*, the CAFC distinguished *Sega* and *Connectix* by saying that they were fair use cases and that compatibility considerations might be relevant to fair use defenses. *Oracle Am., Inc. v. Google Inc.*, 750 F.3d 1339, 1368–71 (Fed. Cir. 2014). However, the Ninth Circuit plainly stated in both cases that interface procedures necessary for interoperability are unprotectable elements of copyrighted programs under § 102(b).

364. See, e.g., Alfred Z. Spector, *Software, Interface and Implementation*, 30 JURIM. J. 79, 85–87, 90 (1989) (discussing distinction between interfaces and implementations).

interoperability. Congress has, moreover, indirectly ratified these court holdings that allow reuse of interfaces necessary for compatibility.³⁶⁵

Java method headers (or declarations, as they are sometimes called) are not lines of code in the sense of executable program instructions.³⁶⁶ They are specifications designed to invoke particular program functions. The method headers must be implemented in program instructions that when compiled, will become executable code. In this respect, the method headers are nonliteral elements of programs, which Google reimplemented in independently written code. The functionality of those method headers is inextricably interconnected with any expression they might be said to contain. The rules of Java, as the District Court noted, constrained Google's design choices as to names of Java methods and classes and as to the structure of the API command structure.³⁶⁷ The merger doctrine was a suitable alternative justification for the holding that the Java APIs at issue in *Oracle* are unprotectable by copyright law, as the District Court held.³⁶⁸

E. THE CAFC ERRED IN INTERPRETING THE MERGER DOCTRINE

The CAFC misinterpreted the merger doctrine in several respects. For one thing, it rejected outright the idea that merger can be a “copyrightability” issue.³⁶⁹ However, merger was in fact treated as a copyrightability issue in *Baker*,³⁷⁰ as well as in numerous other cases.³⁷¹

365. Congress created an exception to rules that forbid bypassing technical protection measures to authorize programmers to reverse engineer technically protected programs to get access to interface information necessary for interoperability. 17 U.S.C. § 1201(f).

366. See, e.g., Mike Masnick, *Yes, The Appeals Court Basically Got Everything Wrong in Deciding APIs Are Covered by Copyright*, TECHDIRT (Aug. 18, 2015), <https://www.techdirt.com/articles/20150817/11362131983/yes-appeals-court-got-basically-everything-wrong-deciding-apis-are-covered-copyright.shtml> [https://perma.cc/9VY6-RN9V] (criticizing the CAFC *Oracle* decision for not understanding what an API is).

367. *Oracle Am., Inc. v. Google Inc.*, 872 F. Supp. 2d 974, 979 (N.D. Cal. 2012) (“[S]ince there is only one way to declare a given method functionality, everyone using that function must write that specific line of code in the same way . . .”).

368. *Id.* at 997.

369. *Oracle Am., Inc. v. Google Inc.*, 750 F.3d 1339, 1354–62 (Fed. Cir. 2014).

370. *Baker v. Selden*, 101 U.S. 99, 103 (1879) (Selden's forms held unprotectable by copyright law as “necessary incidents” to the bookkeeping system).

371. See, e.g., *Ho v. Taflove*, 648 F.3d 489, 497 (7th Cir. 2011) (equation not copyrightable); *ATC Distrib. Group, Inc. v. Whatever It Takes Transmission & Parts, Inc.*, 402 F.3d 700, 707–08 (6th Cir. 2005) (parts numbering system not copyrightable); *Lexmark Int'l, Inc. v. Static Control Components*, 387 F.3d 522, 534–42 (6th Cir. 2004) (printer program not copyrightable); *Warren Publ'g Inc. v. Microdos Data Corp.*, 115 F.3d 1509, 1518 n.27 (11th Cir. 1997) (systematic compilation not copyrightable); *Kern River Gas Co. v. Coastal Corp.*, 899 F.2d 1458, 1463–65 (5th Cir. 1990) (route of gas line not copyrightable). The Copyright Office also recognizes that merger can present a

Second, the CAFC erred in saying that merger can only be found when, *ex ante*, there is just one way to express an idea or function and any creativity in the design of an API makes it protectable by copyright law.³⁷² Other courts, *Altai* among them, have recognized that there sometimes is, *ex ante*, more than one or a few ways to design nonliteral elements of computer programs, but efficiency considerations may narrow the range of options, so that merger applies.³⁷³ The CAFC did not consider whether efficiency considerations might have limited options as to the design of the Java API method headers. Numerous cases have taken other factors into account when assessing merger defenses, such as whether the claimed expression was the most logical and useful way to do something, whether practical considerations or functionality limited options, and whether particular designs were necessary to achieving objectives.³⁷⁴

The CAFC in *Oracle* conjectured that merger might apply to three “core” Java API packages that were necessary to make use of the Java language, which all agreed Google was free to use.³⁷⁵ The design of those core packages was probably no more constrained, *ex ante*, than any others of the API packages. Yet, if the Java language cannot be used at all unless

copyrightability issue. *See* Compendium, *supra* note 143, § 313.3(B) (stating Office will not register claims of copyright when it believes the work consists of merged expression). The CAFC relied upon *Ets-Hokin v. Skyy Spirits, Inc.*, 225 F.3d 1068 (9th Cir. 2000), to say that the Ninth Circuit does not recognize merger as a copyrightability issue. *See Oracle*, 750 F.3d at 1358. While *Ets-Hokin* did reject the merger defense that the photograph at issue was uncopyrightable, this was because the photograph had enough originality to support a copyright. *See* 225 F.3d at 1077, 1082. The scope of that copyright, however, was so “thin” that another photograph of the bottle did not infringe. *See Ets-Hokin v. Skyy Spirits, Inc.*, 323 F.3d 763, 766 (9th Cir. 2003). In *Allen v. Academic Games League of Am.*, the Ninth Circuit treated merger as a copyrightability issue. *See* 89 F.3d 614, 617–18 (9th Cir. 1996); *see also* Samuelson, *Reconceptualizing Merger*, *supra* note 292, at 435–38.

372. *Oracle*, 750 F.3d at 1359–62.

373. *See supra* text accompanying notes 103–106. *ATC* is an example of a non-software copyright case recognizing efficiency as a design constraint. *See ATC*, 402 F.3d at 709.

374. *See, e.g., id.* at 707 (reasonableness as constraint); *Rice v. Fox Broad. Co.*, 330 F.3d 1170, 1177 (9th Cir. 2003) (merger if most logical); *Yankee Candle Co. v. Bridgewater Candle Co.*, 259 F.3d 25, 35 (1st Cir. 2001) (functional considerations); *Matthew Bender & Co. v. West Publ’g*, 158 F.3d 674, 685 (2d Cir. 1998) (feature became standard); *Crume v. Pac. Mut. Life Ins. Co.*, 140 F.2d 182, 184 (7th Cir. 1944) (alternative language might not achieve objective); *Matthew Bender & Co. v. Kluwer Law Book Publishers, Inc.*, 672 F. Supp. 107, 110–11 (S.D.N.Y. 1987) (most logical; practical considerations).

375. *Oracle*, 750 F.3d at 1362. Because of interdependencies among elements of the Java API packages, reuse of at least ten is necessary to implement the Java language specification.

one conforms to the method headers and classes of those three packages, merger would be a reasonable doctrine to apply.

Third, the CAFC incorrectly stated that merger can only be found when the plaintiff (not the defendant) faced constraints in its design decisions.³⁷⁶ But courts often recognize that a defendant's design choices can be constrained by what the plaintiff did.³⁷⁷ To compete effectively and to enable ongoing innovation, a second comer may need to use the same designs. Even CONTU accepted this proposition: "when specific instructions, *even though previously copyrighted*, are the only and essential means of accomplishing a given task, their later use by another will not amount to an infringement."³⁷⁸ This indicates that CONTU accepted that function and expression might merge over time.

Fourth, the CAFC ignored the District Court's finding that there was far less of a distinction between the Java language and the API packages than Oracle acknowledged.³⁷⁹ Insofar as there is little or no difference between the Java language—which all agree is not protectable by copyright law—and the component parts of the API that Google used, this consideration weighs in favor of finding merger in *Oracle*.³⁸⁰

The CAFC should also have taken into account the interests of the nine million Java programmers who have become accustomed to using the Java command structure when expressing themselves in the Java language.³⁸¹ If

376. *Id.* at 1361.

377. *See, e.g.*, *N.Y. Mercantile Exch., Inc. v. Intercontinental Exchange, Inc.*, 497 F.3d 109, 116–19 (2d Cir. 2007); *ATC*, 402 F.3d at 705–09; *Southco, Inc. v. Kanebridge Corp.*, 390 F.3d 276, 282 (3d Cir. 2004); *Lexmark Int'l, Inc. v. Static Control Components*, 387 F.3d 522, 536–42 (6th Cir. 2004); *see also* Samuelson, *Reconceptualizing Merger*, *supra* note 292, at 442–44.

378. CONTU Report, *supra* note 2, at 74 (emphasis added).

379. *Oracle Am., Inc. v. Google Inc.*, 872 F. Supp. 2d 974, 982 (N.D. Cal. 2012).

380. The CAFC also failed to recognize that numerous cases have treated merger as reasonably interchangeable alternatives to § 102(b) defenses. *See, e.g.*, *Ho v. Taflove*, 648 F.3d 489, 497 (7th Cir. 2011); *Warren Publ'g Inc. v. Microdos Data Corp.*, 115 F.3d 1509, 1518 n.27 (11th Cir. 1997); *Gates Rubber Co. v. Bando Chem. Indus.*, 9 F.3d 823, 845 (10th Cir. 1993). As in *Baker*, once an author chooses to create a text embodying an unprotectable system or method of operation, the expressive choices of follow-on creators wanting to use the same system or method are constrained.

381. There is a deep irony in Oracle's copyright lawsuit against Google. Sun Microsystems, whose IP assets Oracle acquired in 2010, was once the foremost proponent of freedom to interoperate, by which it meant there should be no intellectual property protection for APIs insofar as they were necessary to enable interoperability. Sun's Deputy General Counsel Peter M.C. Choy was a lead lawyer on numerous amicus curiae briefs for the American Committee for Interoperable Systems in software copyright cases, including *Altai*, *Sega*, *Bateman*, and *Borland*. These briefs can be found online. *See Interoperability Resources*, COMPUTER & COMMUNICATIONS INDUSTRY ASSOCIATION, <http://www.ccia>

it was appropriate to take into account the interests of users who constructed macros in the Lotus 1-2-3 language in denying Lotus's claim in *Borland*, then it should be appropriate to take into account the third party effects of a ruling in Oracle's favor on programmers accustomed to using the Java command structure that Google decided to include in the Android software and educators who teach them to students. Courts should not be forcing programmers to engage in needless variation when standardization would better accomplish the objectives of copyright law by letting programmers express themselves in the command language they know well.³⁸²

net.org/interop/ [<https://perma.cc/L5QK-BDXP>] (compiling amicus briefs on the interoperability issue).

Consider this excerpt from the ACIS brief to the Supreme Court in support of *Borland*, which echoes arguments that Google made in *Oracle*:

Unlike traditional literary works such as novels and plays that stand alone and do not need to interact with any other work, computer programs never function alone; they function only by interacting with the computer environment in which their developers place them. This environment is absolutely unforgiving. Unless the computer program conforms to the precise rules for interacting with the other elements of the system, no interaction between the program and the system is possible. As a consequence, no matter how much better or cheaper the new program is, it will not enjoy a single sale if it cannot interoperate in its intended environment. If the developer of one part of the environment can use copyright law to prevent other developers from writing programs that conform to the system of rules governing interaction within the environment – interface specifications, in computer parlance – the first developer could gain a patent-like monopoly over the system without ever subjecting it to the rigorous scrutiny of a patent examination. Lotus seeks to use copyright in exactly this manner.

Brief Amici Curiae of American Committee for Interoperable Systems and Computer & Communications Industry Ass'n in Support of Respondent at 4–5, *Lotus Dev. Corp. v. Borland Int'l Inc.*, 516 U.S. 233 (1996) (No. 94-2003), 1995 WL 728487. Oracle cannot perhaps be bound by the legal positions Sun took in those cases, but surely it is fair game to point out the stark contrast between then and now. After all, Google talked to Sun about a license, not Oracle. Moreover, Sun's last CEO testified in support of Google's defense. See Bryan Bishop, *Former Sun CEO Jonathan Schwartz Testifies for Google in Oracle Trial*, THE VERGE (Apr. 26, 2012), <http://www.theverge.com/2012/4/26/2977858/former-sun-ceo-jonathan-schwartz-testifies-for-google-oracle-trial> [<https://perma.cc/8F9R-YH25>].

382. See, e.g., *Lotus Dev. Corp. v. Borland Int'l, Inc.*, 49 F.3d 807, 818 (1st Cir. 1995), *aff'd by an equally divided Court*, 516 U.S. 233 (1996) (stating it would be “absurd” to require users to have to learn new command structures for programs); Burk, *supra* note 206, at 592 (In *Baker* and *Borland*, protecting the forms and commands “would have been tantamount to protecting the method or process they embodied. This is the rationale of merger.”); see also *supra* text accompanying notes 284–286 concerning the difficulties that Java programmers would have encountered if Google had attempted to develop different method headers and classes for a variant version of them for the Android platform.

Command structures and APIs are examples of computer program design elements as to which merger of function and expression may and often does occur. It is important to competition and ongoing innovation in the software industry that merged elements such as these are available for reuse by subsequent programmers, especially given the powerful presence of network effects.³⁸³

V. DIFFERENT CONCEPTUALIZATIONS ON THE RELATIONSHIP BETWEEN PATENT AND COPYRIGHT PROTECTIONS FOR SOFTWARE

Courts have sometimes assigned to copyright law the role of protecting program expression and patent law the role of protecting program processes or other functionality.³⁸⁴ The Ninth Circuit's *Sega* decision, for instance, stated that Sega could not use copyright law to get exclusive rights in the interfaces that constituted the functional requirements for achieving program-to-program compatibility because that kind of protection was available only from patent law.³⁸⁵ The Second Circuit in *Altai* suggested that patents might be more appropriate than copyright for protecting utilitarian nonliteral elements of programs.³⁸⁶ Such statements draw upon the Supreme Court's *Baker* decision, which channeled useful arts to the

383. A prominent economist has noted that it is "inefficient to protect the arbitrary choices whose commercial value is created solely by the network incentives to imitate—and to protect the useful ideas only indirectly by protecting these ancillary innovations. Such protection not only seems likely to have adverse consequences on compatibility, but also protects only indirectly and haphazardly the useful ideas, the costs of whose creation intellectual-property policy is meant to cover." Joseph Farrell, *Standardization and Intellectual Property*, 30 JURIM. J. 35, 49 (1989).

384. *See, e.g.*, *Mitek Holdings, Inc. v. Arce Eng'g Co.*, 89 F.3d 1548, 1556 n.19 (11th Cir. 1996) (copyright protects expression, but not program processes which are the province of patent law); *Gates Rubber Co. v. Bando Chem. Indus.*, 9 F.3d 823, 837 (10th Cir. 1993) (description of program process may be copyrightable, but program process is patentable); *Atari Games Corp. v. Nintendo of Am., Inc.*, 975 F.2d 832, 839 (Fed. Cir. 1992) (copyright protects program expression, and patent law protects program processes and methods). Some commentators have expressed this view as well. *See, e.g.*, Dennis S. Karjala, *The Relative Roles of Patent and Copyright in the Protection of Computer Programs*, 17 J. MARSHALL J. COMP. & INFO. L. 41, 41–42 (1998); Englund, *supra* note 10, at 893–96.

385. *Sega Enters. Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1526 (9th Cir. 1992); *see also* *Apple Computer Inc. v. Microsoft Corp.*, 35 F.3d 1435, 1443 (9th Cir. 1994) ("Apple cannot get patent-like protection for the idea of a graphical user interface . . .").

386. *Comput. Assocs. Int'l, Inc. v. Altai, Inc.*, 982 F.2d 693, 712 (2d Cir. 1992).

patent system and limited the scope of copyright in any work describing or depicting the useful arts to the plaintiff's expression.³⁸⁷

The patentability of a process, system or method of operation embodied or depicted in a copyrighted work should, in keeping with the Supreme Court's ruling in *Baker*, be a factor indicating that that innovation is among the functional design elements that § 102(b) was meant to exclude from copyright protection. It would undermine incentives to use the patent system if innovators could get several times longer duration of IP protection from copyright law without applying for a patent and subjecting their claims to examination for novelty and nonobviousness, among other things.³⁸⁸ Making sure that copyright does not indirectly protect technological innovations that are not, in fact, patented promotes competition and ongoing innovation in the useful arts.

Unfortunately, the functional and expressive elements of computer programs cannot be as readily distinguished as *Baker* and the conventional paradigms of copyright and patent law posit.³⁸⁹ Functionality pervades program design, and creative choices abound, so the usual channeling mechanisms that courts have traditionally applied work less well as applied to computer programs. Courts should develop more subtle ways to conceptualize the relative roles of utility patent and copyright in protecting program innovations than the categorical exclusivity approach that *Baker* and its progeny law suggest.³⁹⁰

Although *Baker* posited that patent and copyright laws protect very different kinds of innovations, Section A explains why courts have not always found categorical exclusivity of intellectual property subject matters to be persuasive and why separating out the roles of patent and copyright in the protection of software innovations has proven to be difficult. Section B

387. *Baker v. Selden*, 101 U.S. 99, 101–03 (1879).

388. *See, e.g.*, *Oracle Am., Inc. v. Google Inc.*, 872 F. Supp. 2d 974, 984, 996 (N.D. Cal. 2012); Karjala, *supra* note 384, at 44–45.

389. *See, e.g.*, Reichman, *supra* note 1, at 2480–81.

390. Inventors of new and useful machines, manufactures, compositions of matter, and processes are eligible to apply for utility patent protection. 35 U.S.C. § 101, et seq. If the patent applicant satisfies the statutory requirements, including articulation of specific claims that will define the scope of the patent, a utility patent will issue. Utility patents have a maximum duration of twenty years from the date of filing. 35 U.S.C. § 154(a)(2). The patentability of computer program innovations has been controversial for decades. *See, e.g.*, Pamela Samuelson, *Benson Revisited: The Case Against Patent Protection for Algorithms and Other Computer Program-Related Inventions*, 39 EMORY L.J. 1025 (1990). The Supreme Court recently struck down software-related patent claims in *Alice Corp. v. CLS Bank*, 573 U.S. ___, 134 S. Ct. 2347 (2014). Patent protection may also be available to creators of ornamental designs for articles of manufacture. 17 U.S.C. § 171, et seq.

explains the flaws in the CAFC's *Oracle* decision's analysis of copyright-patent boundary issues. Section C offers an alternative framework for thinking about those boundaries as applied to software innovations. Section D observes that software developers often rely more on non-IP strategies, such as first mover advantages and complementary assets, to attain competitive advantage than on intellectual property protections.

A. REASONS TO BE CAUTIOUS OF CATEGORICAL EXCLUSIVITY
ARGUMENTS ABOUT PATENT AND COPYRIGHT PROTECTIONS FOR
SOFTWARE INNOVATIONS

There are several reasons why it may be risky for defendants in software copyright cases to argue that a particular nonliteral element of the plaintiff's program (which it copied) cannot qualify for copyright protection because patents have issued for that kind of program innovation.³⁹¹ For one thing, *Baker* notwithstanding, defendant arguments for categorical exclusivity in intellectual property cases have sometimes proven unpersuasive.³⁹² For instance, Mazer lost his argument that Stein's Balinese dancer statuette was ineligible for copyright protection because Stein could have gotten (but did not) a design patent on the statuette for use as a lamp base.³⁹³ The Court was untroubled by the existence of this overlap between design patent and copyright subject matters.³⁹⁴ Stein's statuette was an ornamental design for an article of manufacture, but it also qualified as a copyrightable work of art.³⁹⁵ This explains the Court's dictum that the potential patentability of

391. Google's brief asking the Supreme Court to review the CAFC ruling seemed to make a categorical exclusivity argument. *See* Petition for a Writ of Certiorari at 23–32, *Google Inc. v. Oracle Am., Inc.*, 135 S. Ct. 2887 (2015) (No. 14-410), 2014 WL 5319724.

392. *See* *J.E.M. Ag Supply, Inc. v. Pioneer Hi-Bred Int'l, Inc.*, 534 U.S. 124 (2001) (rejecting argument that novel plants could not be patented because Congress intended for them to be protected only under the Plant Variety Protection Act).

393. *Mazer v. Stein*, 347 U.S. 201, 217 (1954).

394. *See id.* at 214–15. The Court cited to precedents recognizing an overlap of design patent and copyright subject matter. *Id.* at 215 n.33, 217 n.37. Some of these cases had required creators to elect between copyright and design patent protection; the Court chose not to address the election issue. *See id.*

395. There is not much overlap of copyright and design patent subject matters because of copyright's useful article doctrine which excludes PGS works in which expression and functionality have merged. *See supra* notes 297–304 and accompanying text. The integration of functionality and ornamentality does not disqualify designs from design patent protection. *See, e.g.*, Robert C. Denicola, *Applied Art & Industrial Design: A Suggested Approach to Copyright in Useful Articles*, 67 MINN. L. REV. 707, 707–08 (1982) (“Design patents long offered the possibility of protection for the ornamental design of a useful product.”).

that design was irrelevant to whether it was eligible for copyright protection.³⁹⁶

However, neither the Supreme Court nor any other court has recognized subject matter overlap between copyright and utility patent laws.³⁹⁷ No decision has ever held that a creator can get copyright *and* utility patent protection for *exactly* the same aspect of a creation.³⁹⁸ The *Baker* decision suggests that subject matter overlap of these two laws does not exist. Yet, when the Supreme Court was presented with a categorical exclusivity argument in *Borland*, it split evenly on the merits despite *Borland*'s citation to some utility patents on similar innovations as evidence that they were patent, not copyright, subject matter.³⁹⁹

A second reason not to put too much weight on the existence (or not) of utility patent protection for some types of program-related innovations in judging whether copyrights have been infringed is that patents on some innovative designs may have issued at a different level of abstraction than the copyright claim may be alleged to cover.⁴⁰⁰ The District Court in *Oracle* did not, for example, analyze the API patents it mentioned to compare them to the API command structures in which Oracle claimed copyright.⁴⁰¹

Yet, a levels-of-abstraction assessment of the relative roles of patent and copyright in protecting programs may not be a reliable indicator. After all, patent lawyers make strategic decisions in drafting patent applications about the level of abstraction at which to pitch their clients' claims. There are significant advantages to claiming inventions at higher levels of abstraction because if the claim is allowed, it will enable the patentee to enjoy a broader scope of patent protection for the innovation.⁴⁰² Even if patent claims could

396. *Mazer*, 347 U.S. at 216–17.

397. *Mazer* distinguished design patents from utility patents in relation to copyright protections. *Id.* at 215 n.33, 217 (citing approvingly to *Taylor Instrument* for its holding that utility patents and copyrights are mutually exclusive).

398. *See, e.g.*, *Laureyssens v. Idea Group, Inc.*, 964 F.2d 131, 141 (2d Cir. 1992) (patents on puzzle design narrowed scope of copyright in plaintiff's puzzle).

399. *See, e.g.*, Brief for Respondent, *supra* note 200, at 21, 32–34, 43–44. The categorical exclusivity of patent and copyright subject matters has a constitutional character because the Constitution speaks of Congress as having power to grant exclusive rights to authors and to inventors in “their *respective* writings and discoveries.” U.S. CONST., art. I, § 8, cl. 8 (emphasis added).

400. *See, e.g.*, Lemley, *supra* note 17, at 22–27 (discussing the role of software patents and copyrights and levels of abstraction as a way to distinguish their roles in software protection).

401. *Oracle Am., Inc. v. Google Inc.*, 872 F. Supp. 2d 974, 996 (N.D. Cal. 2012).

402. *See, e.g.*, Mark A. Lemley, *Software Patents and the Return of Functional Claiming*, 2013 WISC. L. REV. 905 (2013).

have been drafted to cover lower level functionality, they need not be so limited.

A third reason to be cautious about the patent or copyright subject matter issue in computer program cases is that § 102(b) excludes from the scope of copyright protection more than just patentable procedures, processes, systems, and methods of operation.⁴⁰³ Mathematical innovations are among the fundamental building blocks of knowledge that should be free for reuse as abstract ideas under both copyright and patent law.⁴⁰⁴ Consider, for instance, Benson's algorithms for transforming binary coded decimals to pure binary form, which the Court held was unpatentable as an abstract idea.⁴⁰⁵ That algorithm would unquestionably be part of the SSO of any program that embodied it. If the Benson algorithm is too abstract to qualify for patent protection, it should be a strong candidate for ineligibility for copyright protection under § 102(b) as an abstract mathematical procedure.

A fourth consideration that cuts against categorical exclusivity is that software-related patents might have issued in error.⁴⁰⁶ This is especially pertinent now that the pendulum on the patentability of software has gone from almost-never-available in the 1960s to mid-1980s to almost-always-available in the mid-1980s to the early 2000s.⁴⁰⁷ In the past decade, that

403. See, e.g., *Ho v. Taflove*, 648 F.3d 489, 497 (7th Cir. 2011) (equation held unprotectable by copyright law). Parts numbering systems are similarly ineligible for copyright protection and likely unpatentable as well. See, e.g., *Southco, Inc. v. Kanebridge Corp.*, 390 F.3d 276, 281 (3d Cir. 2004) (parts numbering system unprotectable).

404. *Alice Corp. Pty. Ltd. v. CLS Bank Int'l*, 573 U.S. ___, 134 S. Ct. 2347, 2355 (2014) (patent); *Gates Rubber Co. v. Bando Chem. Indus.*, 9 F.3d 823, 842–43 (10th Cir. 1993) (copyright).

405. *Gottschalk v. Benson*, 409 U.S. 63, 67, 71–72 (1972). The Supreme Court has recently reaffirmed that the Benson algorithm was unpatentable subject matter. See *Alice*, 134 S. Ct. at 2354–55 (citing approvingly to *Benson*).

406. See, e.g., *Guthrie v. Curlett*, 10 F.2d 725 (2d Cir. 1926) (invalidating patent on consolidated tariff index held on novelty grounds). After losing the patent case, Guthrie sued Curlett for copyright infringement. Although Guthrie's copyright was valid, the court held that Curlett only copied non-copyrightable functional elements and therefore had not infringed. See *Guthrie v. Curlett*, 36 F.2d 694 (2d Cir. 1929). The invalid patent was not mentioned in the copyright opinion. The patents to which Borland pointed in its brief to the Supreme Court may also have issued in error. See Brief for Respondent, *Lotus Dev. Corp. v. Borland Int'l Inc.*, 516 U.S. 233 (1996) (No. 94-2003), 1995 WL 728538; see also *supra* note 200 and accompanying text.

407. For a discussion of the pre-1990 software patent case law, see Samuelson, *Benson Revisited*, *supra* note 390, at 1048–1132. For a discussion of the case law on patent subject matter in the 1990s and 2000s, see Michael Risch, *Everything is Patentable*, 75 TENN. L. REV. 591 (2008).

pendulum has swung back to sometimes-available-but-sometimes-not.⁴⁰⁸ During the mid-1980s to the mid-2000s, during the almost-always-available period, the USPTO issued a large number of patents for software innovations.⁴⁰⁹ Quite a few of these have been struck down in the aftermath of the Supreme Court’s decision in *Alice Corp. v. CLS Bank Int’l*, which affirmed the invalidation of patents on software-implemented method and system claims for managing settlement risks for financial transactions.⁴¹⁰

B. THE *ORACLE* DECISION’S ANALYSIS OF COPYRIGHT-PATENT BOUNDARIES WAS FLAWED

Utility patent and copyright laws are, of course, separate laws, and each has a different role to play in protecting intellectual creations, including computer programs. Even if the possibility of patents on certain types of software innovations should not automatically mean that those innovations cannot be copyrighted, competition and innovation policies, as well as freedom of expression policy, should caution against collapsing legal boundaries so that there is substantial or complete overlap in copyright and utility patent subject matters.⁴¹¹

Yet, the CAFC seems to have done just that in response to Google’s API-as-patent-not-copyright-subject-matter argument.⁴¹² It invoked the *Mazer* dictum that “[n]either the Copyright Statute nor any other says that because a thing is patentable it may not be copyrighted,”⁴¹³ seemingly untroubled by the idea of overlapping utility patent and copyright protections for API designs. Had the CAFC read *Mazer* more carefully, however, it would have noticed that in the very next sentence, the Court reaffirmed that utility patents and copyrights were quite different because copyright law cannot be used to protect patentable ideas, only authorial

408. See, e.g., *DDR Holdings LLC v. Hotels.com L.P.*, 773 F.3d 1245, 1259 (Fed. Cir. 2014) (upholding software patent and finding no § 101 subject matter problems). See *infra* note 410 for examples of software patents that have been struck down since *Alice*.

409. See, e.g., Brief Amicus Curiae of IEEE-USA in Support of Grant of Certiorari at 2, *Alice Corp. Pty. Ltd. v. CLS Bank Int’l*, 573 U.S. ___, 134 S. Ct. 2347 (2014) (No. 13-298), 2013 WL 5555082 (nearly one million software patents have issued).

410. See 134 S. Ct. 2347, 2360 (2014). For patents that have been struck down since *Alice*, see, for example, *Content Extraction & Transmission LLC v. Wells Fargo Bank*, 776 F.3d 1343 (Fed. Cir. 2014) (data storage method); *Ultramercial, Inc. v. Hulu LLC*, 772 F.3d 709 (Fed. Cir. 2014) (Internet advertising method); *buySAFE, Inc. v. Google Inc.*, 765 F.3d 1350 (Fed. Cir. 2014) (transaction guaranty method).

411. See, e.g., Julie E. Cohen & Mark A. Lemley, *Patent Scope and Innovation in the Software Industry*, 89 CALIF. L. REV. 1, 27 (2001) (“As patent and copyright law overlap more and more, it becomes critical that they take account of each other.”).

412. *Oracle Am., Inc. v. Google Inc.*, 750 F.3d 1339, 1379–81 (Fed. Cir. 2014).

413. *Id.* at 1380 (quoting *Mazer v. Stein*, 347 U.S. 201, 217 (1954)).

expression.⁴¹⁴ *Mazer* also cited approvingly to *Baker*-inspired precedents holding that, as the Court put it, “the Mechanical Patent Law and Copyright Laws are mutually exclusive.”⁴¹⁵

The CAFC discussed the patent-not-copyright-subject-matter issue at the very end of its *Oracle* decision, characterizing it as one of Google’s “policy-based arguments.”⁴¹⁶ These arguments, the CAFC said, “appear premised on the belief that copyright is not the correct legal ground upon which to protect intellectual property rights in software programs,” as though “patent protection for such programs, with its insistence on nonobviousness and shorter terms of protection, might be more applicable and sufficient.”⁴¹⁷ But that was not what Google was arguing. Its two-fold point was that patents were a more appropriate way than copyright to protect APIs and that copyright could protect the code that implemented an API, but not the API.⁴¹⁸

To support its view that copyright was more suitable for protecting programs than patents, the CAFC cited two journalistic articles.⁴¹⁹ With a rhetorical flourish, the CAFC went on to say that until the Supreme Court or Congress decided to alter IP rules, it felt bound to enforce copyright protection for programs and “declin[e] any invitation to declare that protection of software programs should be the domain of patent law, and only patent law.”⁴²⁰ The CAFC should instead have recognized that “the existence of software patents should make courts less willing to extend the

414. *Mazer*, 347 U.S. at 217 (“Unlike a patent, a copyright gives no exclusive right to the art disclosed.”).

415. *Id.* at 215 n.33 (citing *Taylor Instrument Cos. v. Fawley-Brost Co.*, 139 F.2d 98 (7th Cir. 1943) and *Brown Instrument Co. v. Warner*, 161 F.2d 910 (D.C. Cir. 1947)). The Court also cited two others of *Baker*’s progeny, *Fulmer v. United States*, 103 F. Supp. 1021 (Ct. Cl. 1952) (making parachute did not infringe on copyright in parachute design) and *Muller v. Triborough Bridge Authority*, 43 F. Supp. 298 (S.D.N.Y. 1942) (construction of a bridge did not infringe on copyright in bridge design plan). *Mazer*, 347 U.S. at 217 n.39. These decisions are also more consistent with exclusivity of copyright and patent subject matter than to overlap.

416. *Oracle*, 750 F.3d at 1379–81.

417. *Id.* at 1379–80.

418. Masnick, *supra* note 366.

419. *Oracle*, 750 F.3d at 1380. The CAFC cited to one article published in the *Economist* magazine and another in the *Washington Post* in support of copyright as the better form of protection for software. *Id.* While the CAFC correctly cited to my *CONTU Revisited* article, *supra* note 2, as recommending a sui generis form of protection for programs instead of copyright, it did not cite the most relevant of my articles in which I affirm that copyright protects program code, but should not protect interfaces necessary for interoperability. See Samuelson, *Why Copyright Excludes Systems*, *supra* note 124, at 1962–74.

420. *Id.* at 1381.

coverage of copyright law to ideas and functional elements of programs and more willing to engage in a strict filtration analysis,⁴²¹ especially in cases involving claims of nonliteral infringements.

The CAFC's *Oracle* opinion may reflect that court's anxiety that software would be underprotected by IP law if it ruled in Google's favor so soon after the Supreme Court's *Alice* decision substantially cut back on the availability of patent protection for software-related inventions.⁴²² But the CAFC's *Oracle* decision is at odds not only with *Baker* and a fair reading of *Mazer*, but also with two of the CAFC's prior decisions in software copyright cases as well as other software copyright decisions.⁴²³

C. AN ALTERNATIVE APPROACH TO CONCEPTUALIZING THE ROLES OF
COPYRIGHTS AND PATENTS IN PROTECTING SOFTWARE INNOVATIONS

A more appropriate way to conceptualize the respective roles of utility patent and copyright protection for computer programs may be one akin to the approach the Supreme Court took when presented with an argument about a potential overlap between patent and trademark protection in *TraFFix Devices, Inc. v. Marketing Displays, Inc.*⁴²⁴ TraFFix argued that the existence of an expired patent on a dual spring design to enable roadside signs to withstand and bounce back from big gusts of wind meant that the patented design was ineligible to be protectable as trade dress.⁴²⁵ The Court considered the expired patent to be "strong evidence" that the spring design was too functional to be eligible for trademark protection.⁴²⁶ However, it did not go so far as to rule that the existence of a utility patent for a particular design automatically disqualified it from trade dress protection.⁴²⁷

421. Lemley, *supra* note 17, at 27; *see also* Karjala, *supra* note 384, at 66–69 (arguing that patent protection is more appropriate than copyright for computer program SSO because it is more functional, rather than aesthetic, in nature).

422. Judge O'Malley, who wrote the *Oracle* decision, would have upheld as patentable subject matter all of the claims that Alice made against CLS Bank. *See* CLS Bank Int'l v. Alice Corp., 717 F.3d 1269, 1292–1321 (Fed. Cir. 2010). She joined three of the five pro-patent opinions in *Alice*. Her concerns notwithstanding, the Supreme Court struck down all of Alice's claims. Neither Judge Plager nor Taranto participated in the CAFC's *Alice* decision.

423. *See, e.g., Atari Games*, 975 F.2d at 838–39; *see also Hutchins*, 492 F.3d at 1383–85. *See supra* note 384 for citations to other software copyright decisions distinguishing copyright and patent protection for software.

424. *TraFFix Devices, Inc. v. Marketing Displays, Inc.*, 532 U.S. 23 (2001).

425. *Id.* at 27–28.

426. *Id.* at 29–30. The Court recognized that the very same aspect of the design that MDI claimed as trade dress fell within the scope of the expired patent; it also considered the description of the functional advantages of the design in the patent. *Id.* at 31–32.

427. *Id.* at 35.

Along similar lines, courts in software copyright cases, when presented with evidence that utility patents have issued for the same type of nonliteral element of program design as the plaintiff argues is protectable expression, should consider those patents as relevant evidence about whether the innovation in question is a method or system excluded from copyright protection under § 102(b).⁴²⁸

Consistent with *TrafFix* and numerous software copyright decisions, the District Court pointed to the patents Oracle and its predecessor Sun Microsystems had obtained on some aspects of the Java API as a reason not to extend copyright protection to them,⁴²⁹ asserting that “this trial showcases the distinction between copyright protection and patent protection” for computer program innovations.⁴³⁰ The issue “loom[ed] large, where, as here, the vast majority of the code was not copied and the copyright owner must resort to alleging that the accused stole the ‘structure, sequence, and organization’ of the work.”⁴³¹ The court later noted that software copyright cases decided in recent years had moved away from using “SSO” as a characterization of protectable elements of programs out of “fidelity to Section 102(b) and recognition of the danger of conferring a monopoly by copyright over what Congress expressly warned should be conferred only by patent.”⁴³²

The District Court recognized that copyright owners might try to claim exclusive rights to “a functional system, process or method of operation that belongs in the realm of patents, not copyrights.”⁴³³ This troubled the District Court because patent protection, when available, was of much shorter duration than copyright, and unlike copyrights which provide automatic protection, patents are only available to those who apply and have their claims examined for novelty and nonobviousness, more stringent standards of eligibility than copyright requires.⁴³⁴ To buttress its opinion on this point, the District Court quoted from *Baker* and *Sega* about the dangers of allowing creators to get patent-like monopolies through copyright

428. See, e.g., *Oracle Am., Inc. v. Google Inc.*, 872 F. Supp. 2d 974, 997–98 (N.D. Cal. 2012). Courts might also usefully consider whether the possible ineligibility of program SSO for patenting under *Alice* and other precedents on account of abstractness should mean that this nonliteral element of a program is also ineligible for copyright protection under § 102(b).

429. See *id.* at 996.

430. *Id.* at 984.

431. *Id.*

432. *Id.* at 996.

433. *Id.* at 984.

434. See *id.*

protection.⁴³⁵ It noted that large numbers of patents had issued in recent years for software innovations, and indeed, both Oracle and Sun had gotten patents on some aspects of the Java API (although Oracle did not claim that Google infringed any of them).⁴³⁶

The District Court noted that Oracle made much of the creativity that went into the design of the Java APIs, but this was beside the point.

Inventing a new method to deliver a new output can be creative, even inventive, including the choices of inputs needed and outputs returned But such inventions—at the concept and functionality level—are protectable only under the Patent Act Based on a single implementation, Oracle would bypass this entire patent scheme and claim ownership over any and all ways to carry out methods for 95 years⁴³⁷

The District Court did not consider the patent-not-copyright-subject-matter issue as an independent ground for its ruling. Rather, the consideration merely reinforced the court’s conclusion that the command structure at issue was a system or method of operation excluded from copyright protection under § 102(b). Closer scrutiny of interface patents and a comparison of them with the SSO at issue in *Oracle* might have made that court’s analysis more persuasive. Nonetheless the District Court’s approach to the patent-not-copyright-subject-matter issue was much sounder than the CAFC’s.

D. SOFTWARE DEVELOPERS ATTAIN COMPETITIVE ADVANTAGE BEYOND IP RIGHTS

The CAFC should have been less anxious than they seemingly were in *Oracle* about the receding role of patents and the necessarily thin scope of copyright protection for program innovations because software developers have a multi-faceted, nuanced approach to attaining competitive advantage in the marketplace. A recent empirical study demonstrates that software entrepreneurs consider first mover advantages the most important way to attain advantage.⁴³⁸ Complementary assets (e.g., providing services or customization) are next most important.⁴³⁹ Entrepreneurs rated copyrights, trademarks, and trade secrecy as equally significant in protecting software,

435. *See id.* at 984, 994–96.

436. *Id.* at 996.

437. *Id.* at 998.

438. Stuart J.H. Graham et al., *High Technology Entrepreneurs and the Patent System: Results of the 2008 Berkeley Patent Survey*, 24 BERKELEY TECH. L.J. 1255, 1290 (2009).

439. *Id.*

but only between slightly and moderately important to gaining competitive advantage.⁴⁴⁰ Only a minority of software entrepreneurs owned or were seeking patents.⁴⁴¹ Software patents were rated just over slightly important.⁴⁴² Patents were mainly valued as useful assets for impressing investors.⁴⁴³ It is also important to recognize that software developers are often able to recoup investments through business models that depend very little on intellectual property rights,⁴⁴⁴ such as Google's ad-revenue strategy for recouping its investment in the Android platform.⁴⁴⁵ IP rights play a more modest role in protecting software innovations than many IP lawyers might expect.

VI. REFINING THE TESTS FOR SOFTWARE COPYRIGHT INFRINGEMENT

This Article has analyzed various doctrinal approaches that courts have taken in adjudicating infringement claims in software copyright cases. Courts have generally sought to interpret copyright rules in keeping with traditional principles of that law, but also with an eye to providing sufficient protection for programs to induce investments in their development while leaving breathing room for subsequent programmers to build on what has come before in developing competing or complementary innovations. The outcomes of the software copyright decisions have generally been consistent, even though the doctrinal hooks courts have employed have differed in some respects.⁴⁴⁶

Many disputes have been resolved by applying *Altai's* AFC test, while other cases have relied on the § 102(b) method and process exclusions, the merger doctrine, or fair use. Regardless of which doctrinal approach courts have used, judges have generally taken care to ensure that copyright law should not be interpreted to grant patent-like protection to program innovations. If no single approach to judging software copyright claims has

440. *Id.*

441. *Id.* at 1279. Venture-backed startups were more likely than other software firms to own or seek patents. *Id.*

442. *Id.* at 1292, 1303.

443. *See id.* at 1308–09.

444. *See* Pamela Samuelson, *The Uneasy Case for Software Copyrights Revisited*, 79 GEO. WASH. U. L. REV. 1746, 1776–81 (2011) (giving examples of developments in the software industry, such as the provision of software as a service instead of a product, that lessen the need for copyright protections to attain competitive advantage).

445. Google does not charge money for the installation of Android on mobile devices. *Oracle*, 750 F.3d at 1351.

446. *See, e.g.*, BAND & KATOH, INTERFACES ON TRIAL 2.0, *supra* note 10, at 45–46.

emerged, this is perhaps unsurprising given that copyright doctrines largely evolved to address issues posed by cases involving expressive works of art and literature, not functional processes such as programs. As *Altai* noted, programs are square pegs that courts must try to fit in the round holes of copyright.⁴⁴⁷ But for the Supreme Court's *Baker v. Selden* decision, which established that copyright protects authorial expression, not functionality, courts would be floundering much more than they have.

What we can say with considerable confidence is that *Altai* and its AFC test for software copyright infringement were vast improvements over the *Whelan* framework and test for infringement under which every design decision a programmer made was protectable expression except for the rare elements as to which no alternative choices existed. *Altai* recognized that the utilitarian nature of programs meant that the scope of software copyright protection must necessarily be thin. *Altai* identified several categories of unprotectable elements of programs—efficient designs, externally constrained designs, elements common to works of that kind (i.e., scenes a faire elements) and public domain components—and directed that those elements be filtered out before assessing whether the defendant had infringed a software copyright. The main criticism this Article has levied against *Altai* concerned its failure to mandate, in addition, filtration of § 102(b) methods and processes.

Courts have sometimes applied the § 102(b) exclusions, with or without reference to the AFC test, when plaintiffs have alleged infringement based on copying of unprotectable methods or procedures. Among the elements of programs that have been adjudged unprotectable under § 102(b) are algorithms, program behavior, and command structures needed for achieving interoperability. The merger doctrine has complemented § 102(b) exclusions in numerous cases in which defendants used the necessary incidents to reimplement the same functionality as an existing program. This Article has called for explicit recognition of a merger of function and expression doctrine to supplement the merger of idea and expression and of fact and expression doctrines established in the case law. It has also recognized that copyright and patent law should play different roles in the legal protection of computer programs, even though the boundary lines between these laws, as applied to programs, has proven more elusive to articulate than in respect of other copyright subject matters. The Article concurs in an approach that recognizes that the patentability of some program innovations, as well as the unpatentability of abstract program

447. *Comput. Assocs. Int'l, Inc. v. Altai, Inc.*, 982 F.2d 693, 712 (2d Cir. 1992).

designs, should be taken into account in assessing the proper scope of copyright protection in software.

There is, of course, a simpler way to address software copyright infringement claims. It would begin by recognizing that copyright law can and does protect software developers against market-destructive appropriations of their products through the well-established rule that protects source and object code from illicit copying.⁴⁴⁸ Similarly established is the protectability of audiovisual and other conventionally expressive content that may be displayed when computers are executing program instructions.⁴⁴⁹ Making minor changes to program texts, such as changing variable names, rearranging instructions, or recompiling the code to disguise infringement should also not be tolerated. Translation of programs from one programming language to another or other forms of slavish copying may also qualify as infringement.⁴⁵⁰ The main value of copyright protection for software developers lies in protecting these aspects of programs. If copyright law did nothing more than safeguard these literal and nonliteral elements of programs, the software industry would very likely still thrive. Limiting the scope of software copyrights to these protectable elements would likely make a special test for infringement of these works unnecessary.

Yet, courts have invested so much in articulating and applying tests for software copyright infringement that they may be reluctant to abandon these tests. Although the *Altai* AFC test has thus far been the most stable and widely accepted approach to judging software copyright infringement, several courts have adapted it. Some courts, for instance, have decided that the first step's construction of an abstractions hierarchy for the plaintiff's program is unnecessary when the claim of infringement is based on certain specific elements of the defendant's program.⁴⁵¹ Courts then proceed to consider whether various limiting doctrines of copyright law apply to those elements, in keeping with the *Altai* filtration step. In some cases, application

448. See *supra* text accompanying note 130; see also Weinreb, *supra* note 10, at 1250.

449. See, e.g., *Williams Electronics, Inc. v. Artic Int'l, Inc.*, 685 F.2d 870 (3d Cir. 1982) (videogame graphics protectable).

450. The focus in most software infringement cases has been on the reproduction right, but modifying another firm's software and then selling the modified version to the public would also likely infringe the derivative work right. See, e.g., *Allen-Myland, Inc. v. IBM Corp.*, 746 F. Supp. 520 (E.D. Pa. 1990).

451. See, e.g., *Mitel, Inc. v. Iqtel, Inc.*, 124 F.3d 1366, 1373 (10th Cir. 1997) ("Where, as here, the alleged infringement constitutes the admitted literal copying of a discrete, easily-conceptualized portion of a work, we need not perform complete abstraction-filtration-comparison analysis.").

of the limiting doctrines have made the third step unnecessary, for the plaintiff's claims failed because the doctrines precluded protection for the elements which the plaintiff claimed were her expressions.⁴⁵² Some courts have adapted the *Altai* second step to filter out procedures, processes, methods of operation, and systems.⁴⁵³ Only rarely do courts specify which "golden nuggets" of expression remain after the filtration step, even though *Altai*'s third step had directed these elements of programs to be the starting point of the final step of assessing infringement.

I propose the following refinement of the *Altai* test for software copyright infringement. The first step would require the plaintiff to specify exactly which elements of her program that she alleges as the basis of the infringement claim. In most cases, it will be unnecessary to construct a hierarchy of abstractions for the program as a whole because only certain elements are alleged to infringe. A second step should inquire which, if any, allegedly infringing elements lie outside the scope of copyright protection under various limiting doctrines, such as the exclusion of (a) unoriginal elements; (b) abstract ideas, concepts and principles under § 102(b); (c) facts, know-how, and other public domain elements; (d) common elements for works of that kind, standard programming techniques, and constraints based on market demands under the scenes a faire doctrine; (e) efficient design elements; (f) command structures and APIs necessary for achieving interoperability with other programs or hardware; (g) other instances of merger of function and expression; and (h) procedures, processes, systems, and methods of operation under § 102(b). This would adapt the *Altai* AFC test to make its filtration step more rigorous, excluding a wider range of functional design elements in programs.⁴⁵⁴

452. *Id.* at 1376 ("In sum, although Mitel's values constitute non-arbitrary original expression, they are unprotectable as scenes a faire because they were dictated by external functionality and compatibility requirements of the computer and telecommunications industries.").

453. *See, e.g.*, *Gates Rubber Co. v. Bando Chem. Indus.*, 9 F.3d 823, 842–43 (10th Cir. 1993).

454. Under this broader filtration analysis, Rand Jaslow might not have been held an infringer as to the dental lab program he developed to compete with Whelan. Jaslow certainly infringed Whelan's copyright when selling her program to third parties, but it is less clear, given how the law has evolved, that his competing program infringed. The courts in *Whelan* indicated that Jaslow did not directly translate Whelan's program from one language or another (which might justify the finding of infringement). *Whelan*, 797 F.2d at 1228 ("Dr. Moore testified that although the Dentcom program was not a translation of the Dentalab system, the programs were similar in three significant respects."). Had the AFC test been applied in *Whelan*, the file structures Jaslow used, for instance, may have been efficient for the tasks at hand. The court in *Altai* rejected an overall structural similarity claim on scenes a faire grounds, which might have been true in *Whelan* also. The

In some cases, as in *Lexmark* and *Borland*, there will be no need for a third step because the filtration step results in a judicial conclusion that even if the defendant copied something from the plaintiff's work, the expressive elements of the plaintiff's program were not copied. Yet in other cases, courts should proceed to consider whether the defendant copied a sufficient quantum of expression from the plaintiff's work to be held as an infringer. Adoption of this refined *Altai* test would be consistent with the overwhelming majority of software copyright cases and with longstanding principles of U.S. copyright law. It would also promote competition and ongoing innovation in the software industry, in keeping with the constitutional goal of copyright of promoting progress in science and useful arts.

The CAFC's *Oracle* decision obviously took a very different approach. This Article has shown that the CAFC's *Oracle* misinterpreted *Altai*, misunderstood § 102(b), and misapplied the merger doctrine, as well as major court rulings that have applied these rules. The *Oracle* decision is contrary to Supreme Court rulings and to the CAFC's own precedents, especially in treating utility patent and copyright laws as providing overlapping protections for computer program innovations.⁴⁵⁵ The CAFC also failed to grasp that an API that specifies functions that a program is designed to carry out is fundamentally different from the copyright-protectable program that implements that API in independently written code.

Some may think that the *Oracle* decision, erroneous as it is, will have little impact on subsequent cases. The decision is, after all, an outlier in the case law and involves complicated facts. Software copyright cases will, moreover, generally go to the regional circuits, not to the CAFC, which has no jurisdiction in copyright cases except when there is a patent claim in the case. But *Oracle* has introduced new uncertainties in the law of software

similarities in the operation of certain modules—order entry, invoicing, accounts receivable, end of day procedure, and end of month procedure—appear to be automations of Jaslow's business processes that should have been filtered out under § 102(b). *See id.*

455. The CAFC remanded *Oracle* for retrial of Google's fair use defense. *Oracle Am., Inc. v. Google Inc.*, 750 F.3d 1339, 1372–77 (Fed. Cir. 2014). It concluded that “this is not a case in which the record contains sufficient factual findings on which we could base a de novo assessment of Google's affirmative defense of fair use.” *Id.* at 1377. It also indicated that compatibility considerations could be considered in the fair use context. *Id.* at 1376–77. The CAFC characterized Oracle's argument for granting summary judgment on fair use as “not without force,” but ultimately concluded that a new trial was needed. *Id.* On remand, a jury found Google's reimplementation of the Java APIs constituted fair use. *Oracle Am., Inc. v. Google Inc.*, No. C 10-03561 WHA, 2016 WL 3181206 (N.D. Cal. June 8, 2016), *appeal docketed*, Nos. 17-1118, -1202 (Fed. Cir. Nov. 14, 2016).

copyrights, and software companies now have reasons to litigate to test the CAFC's resurrection of the *Whelan* approach to assessing infringement. Given how many software patents are out there, it may be easy for a plaintiff's lawyer to find one in a client's portfolio or for the client to buy one so that the complaint may allege a defendant infringed patents as well as copyrights.⁴⁵⁶ In such cases, the appeals would go to the CAFC instead of to a regional circuit.

In the aftermath of *Alice* and court decisions striking down software patents, it may be tempting for courts to interpret the scope of copyright protection expansively, as the CAFC did in *Oracle*, because the role of patents in protecting program innovations is receding. Under a kind of conservation of IP incentives theory, copyrights might seem to need to be broader to make up for the fact that patents are providing less protection for program innovations. Copyright does important work in protecting programs, but as the Second Circuit recognized in *Altai*, "fundamental tenets" of this law should not be distorted to fill a perceived gap in legal protection for programs.⁴⁵⁷

If software developers need some additional legal protection for industrial design elements of programs that neither copyright nor patent law can provide, they should take their case to fill that gap to Congress, not use the courts to fill the gap through expansive interpretations of copyright protections.⁴⁵⁸ Sui generis forms of protection for software have been proposed in the past, and perhaps this approach should be reconsidered.⁴⁵⁹ Industrial design laws typically provide a relatively short term of protection against market-destructive appropriations.⁴⁶⁰ At present, however, competition and innovation seem to be thriving in the software industry without additional legal protections. Without substantial evidence to

456. See *supra* note 30 (discussing cases subsequent to *Oracle* in which plaintiff's lawyers have learned this lesson).

457. See *Comput. Assocs. Int'l, Inc. v. Altai, Inc.*, 982 F.2d 693, 712 (2d Cir. 1992) ("While incentive based arguments in favor of broad copyright protection are perhaps attractive from a pure policy perspective, ultimately, they have a corrosive effect on certain fundamental tenets of copyright doctrine." (citation omitted)).

458. See *id.* ("[T]he resolution of this specific issue could benefit from further legislative investigation . . .").

459. See, e.g., Samuelson et al., *Manifesto*, *supra* note 10, 2405–20 (proposing a short term of sui generis protection for the industrial compilations of applied know-how embodied in computer programs).

460. See, e.g., Reichman, *supra* note 1, at 2459–65 (discussing European industrial design laws).

support a change in the law,⁴⁶¹ the courts should continue to apply copyright protections for software in keeping with *Altai*, its progeny, and traditional principles of copyright law.

461. See, e.g., Robert Kastenmeier & Michael Remington, *The Semiconductor Chip Protection Act: A Swamp or Firm Ground?*, 70 MINN. L. REV. 417, 439–42 (1985) (evidence of industry needs, among other things, should inform expansions of IP rules).

PATENT NATIONALLY, INNOVATE LOCALLY

Camilla A. Hrdy[†]

ABSTRACT

Anxiety over the efficacy and fairness of the patent system has spawned a variety of proposals to rely more heavily on direct public subsidies for innovation, such as research grants and tax incentives. Applying economic theories of federalism, this Article shows that these public finance alternatives to IP—which I call “innovation finance”—should sometimes be the responsibility of *subnational* governments such as states and cities, rather than the federal government. The economic theory of federalism prescribes that public goods should be supplied by the smallest jurisdiction that internalizes the costs and benefits of its actions without creating externalities (spillovers) for other jurisdictions. States cannot reliably internalize the benefits of patent regimes that require significant public disclosure of information. But they can internalize many of the economic benefits of direct public spending on innovation. Indeed, a long line of theoretical and empirical research suggests that the economic benefits of innovation—mainly, high-salaried employment and long-term economic growth—remain highly concentrated in certain geographic locations. Therefore, optimal allocation of government requires that we presume subnational jurisdiction over innovation finance *unless* significant cross-border spillovers or some other collective action failure indicates that national intervention is necessary. The result should be more effective innovation policies that are tailored to the needs and conditions of disparate geographic regions and the demands of a mobile populace, and more precise clustering of innovation industries across the country. Moving from theory to reality, the Article demonstrates that states *already* provide significant funding for private sector innovation, and that federal funding for research is actually the exception rather than the rule. Lastly, the Article highlights a major weakness of economic federalism theory in this context: it fails to take into account preexisting inequalities among regions that may prevent under-resourced locations from mobilizing effective innovation strategies, thereby locking them out of the competition to grow innovation clusters. I argue that pervasive regional inequality creates an independent basis for federal intervention.

DOI: <https://dx.doi.org/10.15779/Z38028PC5Z>

© 2016 Camilla A. Hrdy.

[†] Assistant Professor of Law, University of Akron School of Law. Thanks to Jack Balkin, Mario Biagioli, T.J. Chiang, Robert Cooter, John Duffy, David Grewal, Amy Landers, Peter Lee, Greg Mandel, David Marty, Robert Merges, Lisa Larrimore Ouellette, Ben Picozzi, Wendell Pritchett, Joshua Sarnoff, Ted Sichelman, Neil Siegel, David Thaw, Andrew Tutt, Polk Wagner, Daniel Walters, and Christopher Yoo. Thanks to participants at the Yale Law School Innovation Beyond IP Conferences; the Thomas Edison Fellowship roundtables; the 2015 Works in Progress Intellectual Property Colloquium (WIPIP); IP Scholars 2015 at DePaul University School of Law; and the inaugural IP Scholars Roundtable at Texas A&M School of Law. Special thanks to the Yale Law School Information Society Project and the Penn Law Center for Technology, Innovation & Competition for access to support, advice, and conversation. Lastly, thanks to the many editors at BTLJ who worked on this piece, including Joel Broussard, Zachary Flood, and Eric Riedel.

TABLE OF CONTENTS

I.	INTRODUCTION	1303
II.	GEOGRAPHICALLY LOCALIZED BENEFITS OF INNOVATION	1310
A.	CLUSTER THEORY.....	1313
B.	“INNOVATION CLUSTERS” DISTINGUISHED.....	1315
C.	THE BENEFITS OF INNOVATION FOR THE LOCAL ECONOMY.....	1318
D.	WHERE ARE THE INNOVATION CLUSTERS TODAY?.....	1321
III.	INTELLECTUAL PROPERTY AND INNOVATION FINANCE	1322
A.	GOVERNMENT INCENTIVES FOR INNOVATION, GENERALLY	1323
B.	MAJOR DIFFERENCES BETWEEN INTELLECTUAL PROPERTY AND INNOVATION FINANCE.....	1325
1.	<i>Efficiency</i>	1325
2.	<i>Fairness</i>	1328
IV.	THE ROLE OF JURISDICTION IN INNOVATION POLICY CHOICES	1329
A.	COLLECTIVE ACTION FEDERALISM.....	1330
B.	APPLICATION TO INNOVATION POLICY	1332
C.	THE BENEFITS OF LOCAL JURISDICTION IN INNOVATION FINANCE.....	1334
1.	<i>Better Incentives to Design Effective Local Innovation Policies</i>	1334
2.	<i>Tiebout Clustering</i>	1336
3.	<i>Innovation Policy Experiments</i>	1339
D.	LIMITATIONS.....	1341
1.	<i>Positive Externalities</i>	1342
2.	<i>Negative Externalities</i>	1343
3.	<i>Immobility and Broken Political Process</i>	1345
4.	<i>A Note on Geographic Inequality</i>	1346
V.	FEDERAL AND STATE INNOVATION POLICY IN PRACTICE	1347
A.	PATENT LAW.....	1348
1.	<i>From State to Federal Rights</i>	1348
2.	<i>A Contrast with Trade Secrets</i>	1351
B.	INNOVATION FINANCE	1356
1.	<i>U.S. Federal Innovation Finance</i>	1357
2.	<i>U.S. State Innovation Finance</i>	1363
VI.	CONCLUSION	1376

The powers delegated by the proposed Constitution to the federal government are few and defined. Those which are to remain in the State governments are numerous and indefinite. The former will be exercised principally on external objects, as war, peace, negotiation, and foreign commerce The powers reserved to the several States will extend to all the objects which, in the ordinary course of affairs, concern the lives, liberties, and properties of the people, and the internal order, improvement, and prosperity of the State.¹

– James Madison, *The Federalist* No. 45

While competition for innovative technologies and services is increasingly global, the context of innovation—and the benefits it brings in economic growth and high value employment—remains local.²

– National Academy of Sciences, “Growing Innovation Clusters for American Prosperity”

I. INTRODUCTION

The core justification for patent law is the notion that human societies thrive on “innovation”—doing things that are new and in some way better than what existed before—and that people and firms will systematically underinvest in innovation absent incentives.³ But a growing number of academics are dissatisfied with intellectual property (IP) as a solution to the incentives problem, arguing that property rights needlessly raise prices for consumers and hinder future innovation.⁴ As a result, these academics argue that patents should be replaced or supplemented by direct public financing for innovation, such as research grants, tax incentives, or public

1. THE FEDERALIST NO. 45, at 292–93 (James Madison) (Clinton Rossiter ed., 1961).

2. CHARLES W. WESSNER (Rapporteur), NATIONAL RESEARCH COUNCIL OF THE NATIONAL ACADEMIES: SUMMARY OF A SYMPOSIUM, GROWING INNOVATION CLUSTERS FOR AMERICAN PROSPERITY 3 (2011).

3. See *infra* Part II.

4. See generally Mark A. Lemley, *Property, Intellectual Property, and Free Riding*, 83 TEX. L. REV. 1031, 1058–65 (2005) (discussing five separate costs of IP rights: deadweight loss for consumers, reduced incentives to innovate, rent-seeking, costs associated with patent prosecution and litigation, distorted investment in research and development).

venture capital.⁵ I call these public finance alternatives to IP “innovation finance.”⁶

Like patents, innovation finance can serve as a regulatory mechanism to more closely align the private value of innovating with the social value of innovation.⁷ Government already provides direct financing for innovation in many cases, so why not do so more? For example, rather than allowing unrestricted patenting of life-saving drugs, government could directly fund production and permit copying by generic drug manufacturers or subsidize

5. See, e.g., Daniel J. Hemel & Lisa Larrimore Ouellette, *Beyond the Patents—Prizes Debate*, 92 TEX. L. REV. 303, 303–04 (2013) (arguing for a pluralistic innovation policy that incorporates patents, prizes, grants, and tax credits); Camilla A. Hrdy, *Commercialization Awards*, 2015 WIS. L. REV. 13, 13–14 (arguing that commercialization awards can in some instances be a more efficient innovation policy tool than commercialization patents); Amy Kapczynski, *The Cost of Price: Why and How to Get Beyond Intellectual Property Internalism*, 59 UCLA L. REV. 970, 970, 1001–02 (2012) (arguing based on efficiency and distributive justice concerns that government should “pay less attention to IP and more attention to its alternatives” such as government procurement); Peter Lee, *Social Innovation*, 92 WASH. U. L. REV. 1, 47–59 (2015) (arguing that many valuable “social innovations” are supported by non-patent incentives including government grants and social capital markets); see also Brett Frischmann, *Innovation and Institutions: Rethinking the Economics of U.S. Science and Technology Policy*, 24 VT. L. REV. 347, 376–95 (2000) (developing a framework for evaluating and choosing between different innovation incentives including IP, tax, grants, and procurement, and arguing in favor of a “mixed incentives” policy); Joshua D. Sarnoff, *Government Choices in Innovation Funding (With Reference to Climate Change)*, 62 EMORY L.J. 1087, 1116–23 (2013) (providing a “[t]axonomy of [g]overnment [i]nnovation [f]unding [m]echanisms” including tax and other forms of subsidies); Michael J. Graetz & Rachael Doud, *Technological Innovation, International Competition, and the Challenges of International Income Taxation*, 113 COLUM. L. REV. 347, 350 (2013) (noting that government support for R&D “comes in many forms,” including “legal protections for IP” and “tax benefits for both R&D itself and the gains from innovation.”).

6. Public financing for innovation was the primary form of innovation incentive explored at the two Yale Law School “Innovation Beyond IP” conferences held in March 2014 and March 2015, respectively. See also SUZANNE SCOTCHMER, INNOVATION AND INCENTIVES 242–43 (2004) (“[A] single innovation may be funded in two ways: by the public sector out of general revenue, and through proprietary prices under an intellectual property regime.”). There are other ways for government to influence innovation besides IP and innovation finance. For instance, government can use regulations that penalize innovators for failing to innovate. See Ian Ayres & Amy Kapczynski, *Innovation Sticks: The Limited Case for Penalizing Failures to Innovate*, 82 U. CHI. L. REV. 1781, 1781 (2015) (drawing a distinction between penalties for failure to innovate and incentives to innovate, and arguing that under specific circumstances penalties for failure to innovate can play a valuable role in innovation policy); see also Gregory N. Mandel, *Regulating Emerging Technologies*, 1 L., INNOVATION & TECH. 75, 75 (2009) (discussing the challenges of regulating emerging technologies).

7. CHRISTINE GREENHALGH & MARK ROGERS, INNOVATION, INTELLECTUAL PROPERTY, AND ECONOMIC GROWTH 24–25 (2010) (noting that taxes or subsidies can be used to correct negative or positive externalities).

insurance for certain treatments to drive down the price of drugs for consumers.⁸ Rather than relying on patents to foster innovation in nanotechnology, governments could offer more direct funding to firms to increase incentives to enter uncertain nanotechnology markets.⁹

Innovation finance is an enticing prospect with a growing number of supporters from both inside and outside the IP field. For example, a Nobel-prize winning economist has concluded that if government has sufficient information and is able to solve the challenge of raising revenues and discriminating between good and bad research projects,¹⁰ the strategy of directly subsidizing innovation would in every case “dominate that of enhancing intellectual property rights.”¹¹ But this discussion is missing a more fundamental issue. Even if public financing for innovation is sometimes fairer and more efficient than IP, it is not clear whether innovation finance should be supplied or administered by the national government or by *subnational* governments such as states, cities, and metropolitan regions.¹² This Article tackles this question and, more broadly,

8. See Amy Kapczynski & Aaron S. Kesselheim, ‘*Government Patent Use*’: *A Legal Approach to Reducing Drug Spending*, 35 HEALTH AFFAIRS 791, 791 (2016) (arguing government could fund or import generic versions of socially valuable drugs based on its power of sovereign immunity and compensate patent holders under the reasonable compensation mechanism codified in 28 U.S.C. § 1498); Rachel Sachs, *Prizing Insurance: Prescription Drug Insurance as Innovation Incentive*, 30 HARV. J.L. & TECH. (forthcoming 2017), draft available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2767182, at 1 (arguing that prescription drug insurance can be wielded as an innovation incentive to avoid the deadweight loss and distortions created by the patent system).

9. Lisa Larrimore Ouellette, *Nanotechnology and Innovation Policy*, 29 HARV. J.L. & TECH. 33, 36 (2015) (illustrating that in the nanotechnology field governments across the world “have played an essential role not only by funding basic research, but also by crafting infrastructure to lower the barriers to entry, and by providing substantial direct funding to firms to help mitigate the risk of entering uncertain nanotechnology markets”).

10. This is a big “if.” IP scholars fiercely debate whether and when innovation finance should be preferred. See discussion *infra* Part III.

11. See Joseph E. Stiglitz, *Knowledge as a Global Public Good*, in GLOBAL PUBLIC GOODS 308, 312 (Inge Kaul et al. eds., 1999) (concluding that “[i]f government could costlessly raise revenues for financing the support and if government were effective in discriminating between good and bad research projects, clearly this strategy would dominate that of enhancing intellectual property rights, for the latter strategy entails static distortions (the monopoly prices associated with patent rights result in prices exceeding marginal costs) and the inefficient utilization of knowledge”). *But see, e.g.*, F. Scott Kieff, *Property Rights and Property Rules for Commercializing Inventions*, 85 MINN. L. REV. 697, 705–17 (2001) (identifying various problems with “reward alternatives” to patents).

12. This Article uses the term “subnational government” in reference to states, cities, and regional governments; and it sometimes use the term “local government” for convenience. But these levels of government are structurally distinct. States have their own constitutions that delegate authority to local governments within their jurisdictions. SANDRA STEVENSON, UNDERSTANDING LOCAL GOVERNMENT 9 (2009) (“Local

the question of which level of government should be responsible for which aspects of innovation policy. The Article concludes that intellectual property law—specifically patent law¹³—should be federal law, but that innovation finance should, with important exceptions, be supplied by the state, city, or other subnational government in which the innovation actually occurs—that is, where the research is performed and commercialized.

In reaching this conclusion, I rely on what Robert Cooter and Neil Siegel call “collective action federalism theory,”¹⁴ which in turn is derived from longstanding economic theories of federalism.¹⁵ Under collective action federalism theory, federal action is justified for resolving a particular public problem only when subnational action produces external costs or benefits for other jurisdictions (externalities or spillovers), and the costs of

governments are completely beholden to state governments for their existence and authority.”). This Article does not address the debate among local government law scholars regarding how much power local governments should have vis-à-vis states. *See, e.g.*, Nestor M. Davidson, *Cooperative Localism: Federal-Local Collaboration in an Era of State Sovereignty*, 93 VA. L. REV. 959, 962 (2007) (challenging “the prevailing view of local governments as powerless instrumentalities of the states”).

13. This Article does not discuss copyrights and trademarks because the degree to which they operate as innovation incentives is unclear. But the jurisdictional trajectory of state copyrights and trademarks resembles that of patents: they started as state rights and became federal rights as interstate commerce made local protections increasingly infeasible. On remaining state protection for copyright in sound recordings, see Goldstein v. California, 412 U.S. 546, 573 (1973); *see also* Michael Erlinger, Jr., *An Analog Solution in A Digital World: Providing Federal Copyright Protection for Pre-1972 Sound Recordings*, 16 UCLA ENT. L. REV. 45 (2009); Eva E. Subotnik & June Beseck, *Constitutional Obstacles? Reconsidering Copyright Protection for Pre-1972 Sound Recordings*, 37 COLUM. J.L. & ARTS 327, 329–30 (2014). On trademark law’s jurisdictional trajectory from state to federal rights, see Lee Ann W. Lockridge, *Abolishing State Trademark Registrations*, 29 CARDOZO ARTS & ENT. L.J. 597 (2011); MARK MCKENNA, *Trademark Law’s Faux Federalism*, in *INTELLECTUAL PROPERTY AND THE COMMON LAW* (Shyamkrishna Balganesch ed., 2013); Peter S. Menell, *Regulating “Spyware”: The Limitations of State “Laboratories” and the Case for Federal Preemption of State Unfair Competition Laws*, 20 BERKELEY TECH. L.J. 1363, 1381–83 (2005). Section V.C.3 discusses trade secret law briefly, noting that trade secret laws are different from patents because they do not mandate disclosure of information that can easily be replicated in other jurisdictions. Thus, unlike patents, trade secrets can be effectively protected without national jurisdiction.

14. Robert D. Cooter & Neil S. Siegel, *Collective Action Federalism: A General Theory of Article I, Section 8*, 63 STAN. L. REV. 115, 119, 137–38, 183 (2010) (explaining the theory of collective action federalism); *see also* ROBERT D. COOTER, *THE STRATEGIC CONSTITUTION* 105–09 (2000) (discussing national versus local public goods and implications for optimal allocation of governmental authority).

15. Cooter and Siegel acknowledge the economic federalism literature as the major precursor for their theory. *See, e.g.*, Cooter & Siegel, *supra* note 14, at 137 n.102 (noting that an “early formulation” of their approach can be seen in work by economists such as Wallace Oates). For further discussion see *infra* Part IV.

negotiation between subnational governments to resolve these externalities are very high; otherwise, power should be assigned to the smallest unit of government that internalizes the effects of its exercise. Cooter has called this the “internalization principle.”¹⁶ With respect to public financing for public goods, the theory prescribes that the federal government should be responsible for supplying “national public goods” like national security, which produce relatively equal benefits for everyone in the country, but that subnational governments should be responsible for supplying “local public goods” like bridges, parks, and fire protection, which mainly benefit the residents of a particular geographic community.¹⁷

Cooter and Siegel argue that their theory explains and justifies Congress’ power under the IP Clause to “promote the Progress of Science and useful Arts” by “securing” the “exclusive Rights” of “Authors and Inventors” to their “respective Writings and Discoveries.”¹⁸ “Because the problem of unauthorized use extends across state lines,” they write, “the problem is national and Congress is better placed than the states to solve it.”¹⁹ But this Article argues that collective action federalism theory and its guiding internalization principle mandate a different conclusion *depending on which type of innovation policy we select*: intellectual property rights or innovation finance. Although Cooter and Siegel are correct that subnational governments cannot generally internalize the benefits of their patent laws because they cannot efficiently prevent imitation and competition within a national marketplace,²⁰ subnational governments internalize many of the benefits of public spending on innovation.

The common wisdom is that innovation produces significant national and indeed global benefits because knowledge produced in one location inevitably “spills over” to other jurisdictions.²¹ But significant empirical and

16. COOTER, note *supra* note 14, at 137; *see also infra* Part IV.

17. *See infra* Part IV.

18. U.S. CONST. art. I, § 8, cl. 8.

19. Cooter & Siegel, *supra* note 14, at 149.

20. Camilla A. Hrdy, *State Patents as a Solution to Underinvestment in Innovation*, 62 U. KAN. L. REV. 101, 111 (2014) (“Given the externalities associated with the creation of new inventions and the difficulty of protecting them in an interstate market—along with the heavy administrative cost of multiple state patent offices—it would be expensive, inconvenient, and socially wasteful if inventors had to rely *solely* on a patchwork of state rights.”). *C.f.* John F. Duffy, *Harmony and Diversity in Global Patent Law*, 17 BERKELEY TECH. L.J. 685, 694 (2002) (arguing that “[j]ust as the externalities provide a justification for the existence of a patent system, so too do they provide a reason for [global] harmonization [of patent law]”).

21. As David Audretsch and Maryann P. Feldman put it, “there is no reason that knowledge should stop spilling over just because of borders, such as a city limit, state line,

theoretical research shows that the immediate economic impacts of innovation tend to be highly concentrated in the geographic regions in which it occurs.²² These regions are the so-called “innovation clusters” like Silicon Valley, California and Boston, Massachusetts, where high tech firms and their employees reside, consume local services, and pay taxes.²³ According to the economic federalism literature and the internalization principle, public financing for innovation should arguably come from the specific city, state or other subnational community that captures the economic benefits of innovation. The result should be more investment in the kinds of innovation that benefit particular communities, more efficient clustering of mobile firms and residents into different technology areas, and (a fortunate side-benefit of following the internalization principle) more experimentation in law and policy.²⁴

Rather than stopping with this theoretical result, this Article also shows that this conclusion is borne out in practice. Even though states no longer grant patents,²⁵ states supply significant amounts of funding and tax incentives for research and commercialization.²⁶ The federal government certainly plays the dominant role in funding basic science with few commercial prospects and research related to national mission areas such as defense and public health, where subnational governments lack sufficient incentives to supply public financing.²⁷ But the federal government’s support for commercialization and innovation outside of these mission areas is comparably minimal.²⁸ Meanwhile, U.S. states and even some cities are

or national boundary.” See David B. Audretsch & Maryann P. Feldman, *Knowledge Spillovers and the Geography of Innovation*, in 4 HANDBOOK OF URBAN AND REGIONAL ECONOMICS (2004), <http://www.econ.brown.edu/Faculty/henderson/Audretsch-Feldman.pdf>, at 6.

22. See discussion *infra* Section II.C.

23. See discussion *infra* Section II.C.

24. See discussion *infra* Section IV.B.

25. See Camilla A. Hrды, *State Patent Laws in the Age of Laissez Faire*, 28 BERKELEY TECH. L.J. 45, 47 (2013) (“Today patent law is purely a federal creature.”). For a discussion of the extent to which remaining state law incentives for innovation are preempted by federal law, see Jeanne Fromer, *The Intellectual Property Clause’s Preemptive Effect*, in INTELLECTUAL PROPERTY AND THE COMMON LAW (Shyam Baganesh ed., 2013); Camilla A. Hrды, *State Patents as a Solution*, *supra* note 20, at 135–61; Sharon K. Sandeen, *Kewanee Revisited: Returning to First Principles of Intellectual Property Law to Determine the Issue of Federal Preemption*, 12 MARQ. INTELL. PROP. L. REV. 299 (2008).

26. See *infra* Section V.B.2.

27. See *infra* Section V.B.1. On the federal government’s role in basic research funding, see generally JOSH LERNER, *THE ARCHITECTURE OF INNOVATION: THE ECONOMICS OF CREATIVE ORGANIZATIONS* 20–21, 33 (2012).

28. For prior iterations of this observation, see, e.g., LEWIS M. BRANSCOMB & PHILIP E. AUERSWALD, *TAKING TECHNOLOGICAL RISKS: HOW INNOVATORS, EXECUTIVES, AND*

increasingly financing innovation at all phases of development and in a wide range of technology areas.²⁹ As this Article documents, states currently spend billions of dollars a year on a wide variety of innovation initiatives, including grants for research, R&D tax credits, venture financing for technology firms, and educational programs, in the hopes that mobile firms will locate and remain in the region.³⁰

My conclusion—that patent law should be national but that innovation finance often is and often should be subnational—has serious implications for innovation policy and the IP field. Specifically: if we follow the advice of academics who propose relying on public finance rather than patents, at least some of these incentives should be the responsibility of local governments. My conclusion also has implications for the economic federalism and public finance literature. While knowledge is often cited as an example of a “global public good” that creates significant free-rider problems and thus necessitates national if not international intervention,³¹ such statements are contradicted by the reality that, at least within the United States, a not-insignificant amount of funding for knowledge goods comes from local governments. In other words, innovation, and the new knowledge that innovation produces, sometimes behaves more like a local public good than a national public good.

This Article proceeds as follows. Part II discusses theoretical and empirical research suggesting that the immediate economic impacts of

INVESTORS MANAGE HIGH-TECH Risks 144 (2001) [hereinafter Taking Risks] (“Unlike the states, which are politically quite comfortable competing with one another to attract new business through active programs of R&D subsidies, federal politics views with suspicion government programs to assist individual firms.”); *see also* Peter Eisienger, *The Rise of the Entrepreneurial State: State and Local Economic Development Policy in the United States* 241–65 (1988) (discussing states’ commitment to supplying venture capital for small businesses, entrepreneurs, and high-technology enterprises in the 1970s and 80s); Matthew Keller, *The CIA’s Pioneering Role in Public Venture Capital Initiatives*, in *State of Innovation: The U.S. Government’s Role in Technology Technical Development* 110–11 (Fred Block & Matthew Keller eds., 2011) (observing that public venture capital programs aimed at spurring economic development began at the state level and contrasting this to the federal government’s more limited adoption of venture capital initiatives) (citing Eisienger’s work).

29. Maryann P. Feldman & Lauren Lanahan, *State Science Policy Experiments*, in *THE CHANGING FRONTIER: RETHINKING SCIENCE AND INNOVATION POLICY* (Adam B. Jaffe & Benjamin F. Jones eds., 2015) (noting that state expenditures on R&D programs at universities alone are now over \$3 billion and describing states’ increasing expenditures since 1980 on these and other initiatives). For examples of recent state expenditures on science and technology initiatives, see *STATE SCI. & TECH. INST., TRENDS IN TECHNOLOGY-BASED ECONOMIC DEVELOPMENT: LOCAL, STATE AND FEDERAL ACTION IN 2012* 7–9 (2012), <http://ssti.org/sites/default/files/>.

30. *See infra* Section V.B.2.

31. *See* Stiglitz, *Knowledge as a Global Public Good*, *supra* note 11, at 310.

innovation—high-salaried employment, more spending on local businesses, increased tax revenues, and a greater store of “local knowledge spillovers”—are both highly localized and heavily concentrated in certain geographic regions within the country. Part III explains the two most common forms of innovation policy intervention—intellectual property and innovation finance—and the costs and benefits of each. Part IV explicates the collective action federalism theory and situates it within the general theory of economic federalism. Part IV then applies the theory to innovation policy and discusses the theoretical benefits, and limitations, of local innovation finance incentives.

Part V illustrates that this theory appears to be a good descriptor of U.S. innovation policy today. In short: modern U.S. patent law is purely national. But outside basic research and selected mission areas, innovation finance is often subnational. While we lack empirical studies demonstrating whether local innovation programs are worth the cost or whether they are superior to alternative federal programs that might be created, economic federalism theory suggests that this allocation of authority could be good for local economies and good for innovation overall. That said, the theory also highlights serious limitations, including lingering externalities and the pernicious threat of persistent geographic inequality. These are potential areas for federal oversight and/or federal subsidy. Part VI summarizes and concludes.

II. GEOGRAPHICALLY LOCALIZED BENEFITS OF INNOVATION

The common utilitarian conception of the role of patents and of intellectual property rights generally is that IP encourages people and firms to innovate despite the difficulty of appropriating the full value of the benefits conferred by their innovations on society.³² “Innovation”—the application of new ideas to products, processes, or other aspects of a productive enterprise in a way that creates value for creators or

32. R. Polk Wagner, *Information Wants to Be Free: Intellectual Property and the Mythologies of Control*, 103 COLUM. L. REV. 995, 1002 (2003) (“Intellectual property laws . . . allow for the creators of intellectual property to individually capture value associated with the information they present to the world; this is, after all, the fundamental utilitarian bargain, a reward for the creativity or innovation that society wants.”). *But see* Brett M. Frischmann & Mark A. Lemley, *Spillovers*, 107 COLUM. L. REV. 257, 258 (2007) (“[T]here is no reason to think that complete internalization of externalities is necessary to optimize investment incentives . . . Spillovers do not always interfere with incentives to invest; in some cases, spillovers actually drive further innovation.”).

consumers³³—may benefit society at large in many ways. As patent law scholars frequently emphasize, innovation generates new knowledge that in turn produces future innovation.³⁴ Indeed, according to famous studies by Edwin Mansfield, the majority of the benefits of particular innovations have come from developments made by others after the initial adoption.³⁵ In the more immediate term, innovation leads to exciting new products and services (at least for consumers who can afford them),³⁶ greater profits and growth prospects for firms (producer surplus), and potentially quite significant financial savings for consumers (consumer surplus).³⁷ Innovation can also have more pervasive and long-term impacts on human wellbeing. Innovation leads to higher-paying jobs, at least for some kinds of workers,³⁸ and is believed to spur economic growth over time. As Josh

33. Like other recent IP scholarship, this Article departs from the narrow definition of innovation used in patent law; it relies on the definition of innovation adopted by many economists, whose major concern is the impact of innovation on the economy and general wellbeing. See GREENHALGH & ROGERS, *supra* note 7, at 4. For an alternative definition of innovation commonly used in schools of information and communication, see EVERETT ROGERS, *DIFFUSION OF INNOVATIONS* 36 (2003) (“An *innovation* is an idea, practice, or object perceived as new by an individual or other unit of adoption.”).

34. See, e.g., Wagner, *supra* note 32, at 1001–02 (“Creation begets more creation; invention leads to further invention.”).

35. See Edwin Mansfield, *How Rapidly Does New Technology Leak Out?*, 34 J. INDUS. ECON. 217, 217 (1985); see also Frischmann & Lemley, *supra* note 32, at 268 (noting that “[s]tatistical evidence repeatedly demonstrates that innovators capture only a small proportion of the social value of their inventions”).

36. The commercial fruits of innovation may be protected by IP or other forms of market power that limit competition. Some consumers may therefore be priced out. See discussion of deadweight loss *infra* Section III.B.1.

37. Consumer surplus is technically defined as the difference between the amount a buyer is willing to pay for a good and what they actually pay. See GREGORY MANKIW, *PRINCIPLES OF MICROECONOMICS* 137 (2010). For example, if an innovation in production allows a company to reduce the costs of producing a popular product and leads the company to lower prices, some consumers will pay less for the product than they can afford to pay. See GREENHALGH & ROGERS, *supra* note 7, at 12–14 (discussing the effects of innovation on consumers); Frischmann & Lemley, *supra* note 32, at 268 (discussing consumer surplus as a form of spillover generated by innovation). But importantly, innovation does not necessarily translate into consumer surplus, especially if firms have significant market power. For instance, through intellectual property rights, lead-time advantage, or secrecy, firms can keep prices high and experience the innovation’s value entirely through profits (producer surplus). SCOTCHMER, *supra* note 6, at 263 (discussing the relationship between, economic growth, consumer surplus, and IP); see also *supra* MANKIW, at 141 (defining producer surplus as the amount a seller is paid for a good minus its costs of production).

38. On the complex relationship between technological innovation and employment, see ERIK BRYNJOLFSSON & ANDREW MCAFEE, *RACE AGAINST THE MACHINE* 36–52 (2011). See also, e.g., Vincent Van Roy, Daniel Vertesy, & Marco Vivarelli, *The Job-Creation Effect of Patents: Some Evidence from European Microdata* (working paper, Apr.

Lerner puts it, “[i]nnumerable studies have documented the strong connection between new discoveries and economic prosperity across nations and over time. The relationship is particularly strong in advanced nations—that is, countries that cannot rely on copying others or on a rapidly increasing population to spur growth.”³⁹

But the economic impacts of innovation on society are not equal across geographic regions. To the contrary, a long line of theoretical and empirical research suggests that, even if innovation ultimately produces national and even global economic benefits, innovation’s major economic impacts are often highly localized. Even within a nation, *even within a state*, many of the concrete economic benefits of technological advance—new sources of profits and tax revenues, new forms of employment, and higher worker wages—are “overwhelmingly concentrated in a small number of geographic locations.”⁴⁰ These lucky winners are the “innovation clusters”: regional economies made up of a critical mass of firms, institutions, and highly skilled talent whose core activities involve high levels of innovation.⁴¹

2016) (finding that the positive impact of employment is statistically significant for firms in the high-tech manufacturing sector, but not significant in low-tech manufacturing and services), <http://ssrn.com/abstract=2770525> (last visited May 22, 2016), at 1.

39. LERNER, ARCHITECTURE OF INNOVATION, *supra* note 27, at 16 (“Since the pioneering work of Moses Abramowitz and Robert Solow in the 1950s, we have understood that technological change is critical to economic growth: innovation has not just made our lives more comfortable and longer than those of our great-grandparents, but has made us richer as well.”); *see also* Graetz & Doud, *supra* note 5, at 348 (noting that it is “clear and essentially uncontested among economists” that technological innovation is important to economic growth). For further discussion of the impacts of innovation on economic growth at the national level, *see*, for example, William Hubbard, *Competitive Patent Law*, 65 FLA. L. REV. 341, 349–52 (2013) (discussing the impact of domestic innovation on the United States’ competitiveness in the global economy).

40. *See* ENRICO MORETTI, THE NEW GEOGRAPHY OF JOBS 73–120 (2012) (documenting the divergent impact of the innovation sector on geographically distinct regions in the United States). As Moretti puts it,

[i]nnovation creates enormous social benefits, in the form of new drugs, better ways to communicate and share information, and a cleaner environment. These benefits are diffuse, in the sense that consumers all over the world can enjoy them. But innovation also creates benefits in the form of new and better jobs. These benefits are overwhelmingly concentrated in a small number of geographic locations.

Id. at 81; *see also*, e.g., David Ibrahim, *Financing the Next Silicon Valley*, 87 WASH. U. L. REV. 717, 719 (2010) (noting that “[h]igh-tech firms are important drivers of U.S. economic growth in today’s knowledge economy, but gains from innovation-based economic growth are highly skewed toward a few regions”).

41. WESSNER, *supra* note 2, at 3 (“Innovation clusters are regional concentrations of large and small companies that develop creative products and services, along with

A. CLUSTER THEORY

Generally speaking, clusters are “geographic concentrations of interconnected companies and institutions in a particular field.”⁴² The core tenet of cluster theory is that when participants in a field, including competitors, workers, and related firms and institutions, locate in the same physical space, they benefit from one another’s presence.⁴³ In 1994, Michael Porter documented a number of industry “clusters” across the United States in a variety of industries, such as microelectronics, biotechnology, aircraft design, casinos, sawmills, clocks, agricultural equipment, and specialty foods.⁴⁴

Although Porter presented the clustering phenomenon as novel—“a kind of new spatial organizational form in between arm’s-length markets on the one hand and . . . vertical integration, on the other”⁴⁵—cluster theory is based on the longstanding concept of “agglomeration benefits.”⁴⁶ The idea

specialized suppliers, service providers, universities, and associated institutions. Ideally, they bring together a critical mass of skills and talent and are characterized by a high level of interaction among these entrepreneurs, researchers, and innovators.”); *see also* MARK MURO & BRUCE KATZ, BROOKINGS, *THE NEW ‘CLUSTER MOMENT’: HOW REGIONAL INNOVATION CLUSTERS CAN FOSTER THE NEXT ECONOMY* 11 (2010) (“Regional innovation (or industry) clusters are geographic concentrations of interconnected businesses, suppliers, service providers, coordinating intermediaries, and associated institutions like universities or community colleges in a particular field (e.g., information technology in Seattle, aircraft in Wichita, and advanced materials in Northeast Ohio.”), <http://www.brookings.edu/research/papers/2010/09/21-clusters-muro-katz>. Other terms used to describe regional economies made up of communities whose core activity involves high levels of innovation include “innovation hubs” and “brain belts.” *See* MORETTI, *supra* note 40, at 82–88 (using “innovation hubs”); ANTOINE VAN AGTMAEL & FRED BAKKER, *THE SMARTEST PLACES ON EARTH WHY RUSTBELTS ARE THE EMERGING HOTSPOTS OF GLOBAL INNOVATION* 1–21 (2016) (using “brain belts”).

42. Michael E. Porter, *Clusters and the New Economics of Competition*, HARV. BUS. REV. (Nov.–Dec. 1998), at 78.

43. *Id.* at 81 (asserting that “[b]eing part of a cluster allows companies to operate more productively in sourcing inputs; accessing in formation, technology, and needed institutions; coordinating with related companies; and measuring and motivating improvement”). For a critical view of cluster theory and Porter’s work, in particular, see Gilles Duranton, *California Dreamin’: The Feeble Case for Cluster Policies*, 3 REV. ECON. ANAL. 3, 3–4 (2011) (arguing that much of the literature on cluster theory lacks theoretical and empirical rigor).

44. Porter, *supra* note 42, at 82.

45. *Id.* at 79.

46. *See* PAUL KRUGMAN, *GEOGRAPHY AND TRADE* 35–67 (1991) (discussing agglomeration benefits and the phenomenon of localization of industry). On agglomeration benefits generally, *see* Lee Anne Fennell, *Agglomerama*, 2014 BYU L. REV. 1373, 1378–79 (2014); BRENDEN O’FLAHERTY, *CITY ECONOMICS* 16–23 (2005); Daniel B. Rodriguez & David Schleicher, *The Location Market*, 19 GEO. MASON L. REV. 637, 639 (2012).

is that firms or individuals operating in a particular industry or trade, be it high technology or shoe-making, benefit when they locate near-by to one another because they can draw on the same markets of specialized labor, the same specialized suppliers and other infrastructure, and can more freely engage in exchange of ideas.⁴⁷ As economist Alfred Marshall put it, observing industry localization in the late nineteenth century, when people in the same trade locate near-by, “the mysteries of the trade become no mystery; but are as it were in the air”⁴⁸

One important result of cluster theory is increased productivity for all members of the cluster and—crucially for this Article—superior capacity to innovate due to locational proximity to others engaged in similar endeavors.⁴⁹ Another important result is that members of a particular field should tend to locate in regions in which others in that field are *already located*.⁵⁰ On this view, “the presence of a large number of firms and workers acts as an incentive for still more firms and workers to congregate at a particular location.”⁵¹ A corollary is that regions in which members of a field are already located *can expect more to follow*.⁵² So, for instance, if a single manufacturing company locates in a city, we should expect others companies that make the same product, as well as workers and suppliers, to locate in the same place and thereby benefit from sharing resources, talent, and ideas.⁵³ This expansion of cluster size should theoretically continue

47. See KRUGMAN, *supra* note 46, at 36–38 (quoting and discussing Marshall’s classic analysis).

48. *Id.* at 37 (quoting Marshall).

49. See Porter, *supra* note 42, at 81–84 (discussing various benefits that result from co-location); *id.* at 83 (“In addition to enhancing productivity, clusters play a vital role in a company’s ongoing ability to innovate.”).

50. In work that won him a Nobel Prize, Paul Krugman made the case that “increasing returns to scale are in fact a pervasive influence on the economy, and that these increasing returns give a decisive role to history in determining the geography of real economies.” See KRUGMAN, *supra* note 46, at 10.

51. *Id.* at 66–67.

52. See, e.g., Timothy Bresnahan, Alfonso Gambardella, & AnnaLee Saxenian, ‘Old Economy’ Inputs for ‘New Economy’ Outcomes: Cluster Formation in the New Silicon Valleys, in CLUSTERS, NETWORKS, AND INNOVATION 116 (Stefano Breschi & Franco Malera eds., 2005) (arguing that people and firms in the innovation sector choose to locate in regions “where other technology firms are already located” to obtain privileged access to their know-how).

53. See Michael Greenstone, Richard Hornbeck & Enrico Moretti, *Identifying Agglomeration Spillovers: Evidence from Winners and Losers of Large Plant Openings*, 118 J. POL. ECON. 536, 536 (2010) (estimating the impact of the opening of a large manufacturing plant for incumbent plants in the same county).

until the benefits of co-location are outweighed by the costs, like traffic, crowding, pollution, and high prices for housing.⁵⁴

B. “INNOVATION CLUSTERS” DISTINGUISHED

Over the years many researchers have focused their attention “on the phenomenon of clusters of *innovation* as distinct from clusters of production activities.”⁵⁵ These days, it is far more common to see scholarship assessing the role of proximity and agglomeration benefits in high technology clusters like Silicon Valley than in clusters devoted to, say, paper manufacturing.⁵⁶

54. Mario A. Maggioni, *Mors tua, vita mea? The Rise and Fall of Innovative Industrial Clusters*, in CLUSTER GENESIS: TECHNOLOGY-BASED INDUSTRIAL DEVELOPMENT 220–23 (Pontus Braunerhjelm & Maryann P. Feldman eds., 2006) (discussing the effect of agglomeration benefits and agglomeration costs on cluster size).

55. See Luigi Orsenigo, *Clusters and Clustering: Stylized Facts, Issues, and Theories*, in CLUSTER GENESIS 196, 195–218 (Pontus Braunerhjelm & Maryann P. Feldman eds., 2006). For a discussion of cluster theory, major concepts, and contributions from diverse economic fields, see Stefano Breschi & Franco Malerba, *Clusters, Networks, and Innovation: Research Results and New Directions*, in CLUSTERS, NETWORKS AND INNOVATION 1–5 (Stefano Breschi & Franco Malera eds., 2005); see also Maryann P. Feldman & Dieter Kogler, *Stylized Facts in the Geography of Innovation*, in HANDBOOK OF THE ECONOMICS OF INNOVATION 381, 383–90 (Bronwyn H. Hall & Nathan Rosenberg eds., 2010) (reviewing literature on the spatial concentration of innovation activity in certain regions).

56. For example, there is a vibrant debate in legal scholarship regarding what caused the success of regions like Silicon Valley. Some argue that locational proximity and open flows of people and information contribute to regions’ success. See ANALEE SAXENIAN, REGIONAL ADVANTAGE: CULTURE AND COMPETITION IN SILICON VALLEY AND ROUTE 128 4 (2d ed. 2006) (discussing different “regional network-based industrial systems” and asserting that “[n]etwork systems flourish in regional agglomerations where repeated interaction builds shared identities and mutual trust while at the same time intensifying competitive rivalries”); see also ORLY LOBEL, TALENT WANTS TO BE FREE: WHY WE SHOULD LEARN TO LOVE LEAKS, RAIDS AND FREE RIDING 76–79 (2013) (discussing the role of agglomeration benefits in high-tech economies such as Silicon Valley and asserting that “[s]uccessful regions depend on a population of skilled and talented workers, and in turn these workers learn more quickly when they work in successful areas”). Some argue that law has played a significant role. See, e.g., Anupam Chander, *How Law Made Silicon Valley*, 63 EMORY L.J. 639, 641–42 (2014) (arguing that while standard accounts assert that Silicon Valley’s success can be explained by the economies of agglomeration, law—and IP and privacy laws in particular—played a more significant role in Silicon Valley’s rise and its global success than has been previously understood); LOBEL, *supra* note 56, at 75 (asserting that “[t]he research points strongly to the many benefits of weaker noncompetes,” and that “more competition and less control of talent flow encourage job growth, start-ups, and regional development”). Others disagree with this conclusion. See Jonathan M. Barnett & Ted Sichelman, *Revisiting Labor Mobility in Innovation Markets*, working paper (2016) (critically examining the evidence behind the assumption that legal regimes that enforce contractual and other limitations on labor mobility deter technological innovation), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2758854 (last visited Nov. 3, 2016).

So what makes innovation clusters different? At first glance, the agglomeration benefits that drive localization in high-tech clusters like Silicon Valley appear quite similar to those that drive localization in industries that engage in little research.⁵⁷ The same factors—sharing inputs like suppliers and infrastructure and the ability to draw on the same pools of specialized labor—might explain location decisions and resulting increases in productivity in either case. As Krugman puts it, discussing the Silicon Valley and Route 128 success stories, “[t]his is just the labor pooling story; the fact that the skill involves high technology, rather than shoemaking or tufting, may be of secondary importance.”⁵⁸

But some argue innovation benefits more from proximity than rote production activities, making co-locating even more important and magnifying the benefits of co-location for all members of a cluster.⁵⁹ The reason for this is said to be linked to Marshall’s observation that locational proximity gives those in the area—and only those in the area—privileged access to knowledge and information that is, so to speak, “in the air.”⁶⁰ Since Marshall’s work in the late nineteenth century, many scholars have studied the particular role in regional economies of what are now commonly called “local knowledge spillovers”: defined as “ ‘knowledge externalities bounded in space,’ which allow companies operating nearby important knowledge sources to introduce innovations at a faster rate than rival firms located elsewhere.”⁶¹ Local knowledge spillovers can encompass a broad range of knowledge and information, including “tacit” know-how required to practice science-based inventions,⁶² informal “know-how trading”

57. See KRUGMAN, *supra* note 46, at 63–67.

58. See KRUGMAN, *supra* note 46, at 65.

59. See, e.g., Luigi Orsenigo, *Clusters and Clustering*, in CLUSTER GENESIS: TECHNOLOGY BASED INDUSTRIAL DEVELOPMENT 196 (Pontus Braunerhjelm & Maryann P. Feldman eds., 2006) (discussing research in the geography of innovation that suggests “innovation is more spatially concentrated than production activities”); see also Feldman & Kogler, *supra* note 55, at 385 (“Innovation is more geographically concentrated than production. Even after controlling for the geographic distribution of production, innovation exhibits a pronounced tendency to cluster spatially.”).

60. KRUGMAN, *supra* note 46, at 37 (quoting Marshall).

61. See, e.g., Stefano Breschi & Francesco Lissoni, *Knowledge Spillovers and Local Innovation Systems: A Critical Survey*, Liuc Papers n. 84, SERIE ECONOMIA E IMPRESA (Mar. 2001), at 1, <http://www.biblio.liuc.it/liucpap/pdf/84.pdf>. For a literature review, see *id.* at 1–10. See also GREENHALGH & ROGERS, *supra* note 7, at 204–05 (discussing various studies suggesting knowledge flows occur more rapidly in proximity despite improvements in distance communication).

62. MICHAEL POLYANI, THE TACIT DIMENSION (1966) (discussing science-based knowledge that can only be learned through personal exchange and practice).

between employees at different firms located in the same area,⁶³ and—most broadly—what Eric Von Hippel calls “sticky information.”⁶⁴

Unlike Kenneth Arrow, who assumed that “[t]he cost of transmitting a given body of information is frequently very low,” Von Hippel argues that some information is simply “sticky,” meaning it cannot easily be transmitted or cannot be transmitted without incurring significant costs.⁶⁵ The degree of stickiness depends on various factors, such as the geographic significance of the innovation and how much tacit science-based knowledge is involved that cannot be easily be codified.⁶⁶ The point is that when the cost “to acquire, transfer, and use information” is high, then it may be more cost-effective to collocate the problem-solvers (people) and the resources required to solve the problem, such as factories, equipment, and natural resources, in a single location.⁶⁷ Some hypothesize that the stickiness of *highly technical* information, in particular, explains why companies wishing to license inventions from universities tend to locate nearby to the university,⁶⁸ why corporate labs are often designed to bring many experts together into one place,⁶⁹ and why firms in fast-paced fields tend to locate nearby to their direct competitors.⁷⁰

63. Informal know-how trading refers to accumulated practical skills and proprietary information routinely and informally traded between innovators at different firms, even rivals. Von Hippel has conducted a case study of process know-how trading in the U.S. steel industry. See ERIC VON HIPPEL, *THE SOURCES OF INNOVATION* 6, 76–77 (1988).

64. Eric von Hippel, “Sticky Information” and the Locus of Problem Solving Implications for Innovation, 40 *MGMT. SCI.* 429–39 (1994), available at <http://web.mit.edu/people/evhippel/papers/stickyinfo.pdf>; see also David Teece, *The Strategic management of technology and intellectual property*, in *COMPETING THROUGH INNOVATION: TECHNOLOGY STRATEGY AND ANTITRUST POLICIES* 3–30, 5–9 (David Teece ed., 2013) (discussing the relationship between the transferability of information and companies’ ability to exclude and appropriate the value of information).

65. *Id.* at 429 (quoting Kenneth Arrow, *Economic Welfare and the Allocation of Resources for Invention*, in *THE RATE AND DIRECTION OF INVENTIVE ACTIVITY* 609, 614–15 (1962)).

66. *Id.* at 429–39.

67. *Id.* at 429.

68. See Peter Lee, *Transcending the Tacit Dimension: Patents, Relationships, and Organizational Integration in Technology Transfer*, 100 *CALIF. L. REV.* 1503, 1536–40 (2012) (discussing the importance of geographic proximity in licensing technology generated at universities and citing empirical studies purporting to confirm the role of proximity in capturing university knowledge spillovers).

69. LERNER, *THE ARCHITECTURE OF INNOVATION*, *supra* note 27, at 21–23 (explaining how the need to combine many experts and the importance of proximity for encouraging knowledge flows can affect corporate structure of corporate research labs).

70. Porter, *supra* note 42, at 83 (arguing that clustering plays a vital role in companies’ “ongoing ability to innovate” because companies within clusters can obtain important information more quickly).

C. THE BENEFITS OF INNOVATION FOR THE LOCAL ECONOMY

There is also reason to believe that the economic gains for the surrounding community are more extensive in the “innovation sector” than in other industries.⁷¹ There are three main reasons for this. First, as just explained, innovation is thought to benefit more from locational proximity than non-innovative production. This means that high-tech firms tend to be located nearby to one another, and that once high-tech companies cluster in an area, others are likely to follow.⁷²

Second, companies in the innovation sector tend to make higher profits because of the way they make their money.⁷³ By definition, an innovation must be to some degree *novel*.⁷⁴ This novelty feature is significant, economically speaking, because—while it may require significant up-front investment—novelty, once achieved, allows innovators to command market power and therefore to charge above marginal cost (more than it costs to produce one additional unit of a good or service.)⁷⁵ This means higher profits, which can be distributed to owners or funneled into further research or employee wages.⁷⁶ Higher profits also means innovators will likely pay more in taxes.⁷⁷ As a result, innovative industries’ impact on producers,

71. The following section distills many of the points made by Enrico Moretti in the second chapter of his 2012 book, *The New Geography of Jobs*. See MORETTI, *supra* note 40, at 45–72. For Moretti, defining features of the “innovation sector” include that firms in the sector rely on innovation to make their profits; the innovation sector firms often perform significant research up front; and they tend to “make intensive use of human capital and human ingenuity.” *Id.* at 55, 67, 48.

72. MORETTI, *supra* note 40, at 62 (“This clustering effect also exists in manufacturing, but it is particularly strong in high tech . . .”).

73. *Id.* at 67 (“Innovation industries are fundamentally different from all other industries in how they make their profits.”).

74. An innovation need not be universally novel. It need only be new to a firm and new to some geographic market. See GREENHALGH & ROGERS, *supra* note 7, at 5 (“We define an innovation as new to the firm and new to the relevant market.”).

75. See GREENHALGH & ROGERS, *supra* note 7, at 9–12 (discussing the effect of innovation on a producer’s ability to charge above marginal cost); see also MANKIW, *supra* note 37, at 268, 303–08 (defining marginal cost as the increase in total cost that arises from an extra unit of production, and explaining that in competitive markets price equals marginal cost, but that in monopolized markets price exceeds marginal cost).

76. See GREENHALGH & ROGERS, *supra* note 7, at 11–12 (explaining that when a monopolist is threatened with entry, an innovation can lower prices and benefit consumers while also leading to higher profits). Obviously, when the innovation is protected by a patent this effect is magnified. See MANKIW, *supra* note 37, at 309–10 (explaining that a patent enables the manufacturer of a pharmaceutical drug to charge above marginal cost and thereby increase profits until the patent term runs out).

77. That said, in practice many nations adopt tax breaks for IP owners that locate intangible assets in the jurisdiction rather than taxing them. For example, the increasingly

consumers, workers, and the overall economy can be far greater than in non-innovative industries.⁷⁸

Third, innovative firms tend to hire high-skilled workers that are in lower supply and thus in high demand—meaning they pay higher wages.⁷⁹ As Erik Brynjolfsson and Andrew McAfee explain, college educated workers have seen significant gains in their salaries over the past four decades, while wages for the less educated have stagnated.⁸⁰ The reason, they argue, is that while low-skill jobs such as traditional manufacturing can be delegated to machines or to humans receiving very low wages, more complex and creative operations like science, programming, management, or marketing decisions “remain the purview of humans.”⁸¹ The more educated, skilled, and creative humans are, the higher they are in demand within innovation-intensive industries, and the more attractive their salaries must be.⁸² In other words, technological advance increases the price (wages) of high-skilled workers as compared to low-skilled workers.⁸³

In new research, economist Enrico Moretti has shown that people living in “brain hubs”—metropolitan areas with higher shares of college-educated workers and often higher shares of patents—do in fact have higher salaries.⁸⁴ Specifically, “college graduates in brain hubs make between \$70,000 and \$80,000 a year, or about 50% more than college graduates in the bottom group.”⁸⁵ Importantly, these gains are not restricted to innovator firms and high-skill employees. According to Moretti, innovation comes with a strong “multiplier effect”: economic gains for everyone located in

common “patent box” offers a preferential tax rate for patent income. *See* Graetz & Doud, *supra* note 5, at 362–75 (discussing patent boxes in Europe and a proposed patent box in the United States). Also, IP owners may manage to avoid paying taxes on their IP by shifting their intangible assets to foreign jurisdictions. *Id.* at 399–401 (noting that companies use income-shifting techniques to “deflect IP income to low- or zero-tax countries even in circumstances where the value of the IP was created in the United States and the resulting products are sold in the United States”).

78. *See* MORETTI, *supra* note 40, at 47–72 (explaining many reasons why innovative industries tend to have greater impacts on the economy).

79. *See* MORETTI, *supra* note 40, at 72.

80. BRYNJOLFSSON & MCAFEE, *supra* note 38, at 39–40.

81. *Id.* at 39.

82. *Id.* at 40; *see also* MORETTI, *supra* note 40, at 72 (“The supply of skilled and creative workers capable of innovating is increasing worldwide, as a growing number of young people in emerging economies obtain college and postgraduate education. But the demand for skilled and creative workers is rising even faster.”).

83. BRYNJOLFSSON & MCAFEE, *supra* note 38, at 40.

84. MORETTI, *supra* note 40, at 88–97.

85. *Id.* at 93, 94–95; *see infra* Section II.D (discussing these brain hubs in more detail).

the surrounding economy.⁸⁶ When innovation industries become established in a location, this increases economic activity, employment, and even salaries for those who provide (non-innovative) local services like restaurants, barber shops, and retail.⁸⁷

Moretti argues this multiplier effect is stronger in the innovation sector, mainly because workers in the innovation sector have more disposable income to spend on local services and more money to spend on construction and real estate.⁸⁸ In addition, he argues that high-skilled workers *spread knowledge and skills* to others in the area, increasing their earning power. Somewhat amazingly, Moretti shows that in places where high-skilled workers reside, there is a positive correlation between the number of skilled workers in the city and the salary of their *unskilled* neighbors.⁸⁹ Specifically, “the earnings of a worker with a high school education rise by about 7 percent as the share of college graduates in his city increases by 10 percent.”⁹⁰ Moretti hypothesizes that skills that pass from workers operating in close proximity to one another—a phenomenon he calls “human capital externalities.”⁹¹

The upshot is that the economic gains from innovation for the overall regional economy are extensive, and they are greater than in non-innovation sectors like traditional manufacturing.⁹² For these reasons, Moretti concludes, “More than any other sector, innovation has the power to reshape the economic fates of entire communities, as well as their cultures, urban

86. MORETTI, *supra* note 40, at 13, 55–63 (discussing research suggesting that attracting scientists, software engineers, and other high-skill workers to a region increase demand for local services such as restaurants, hairdressers, therapists, and yoga instructors).

87. MORETTI, *supra* note 40, at 55–63, 97–101; *see also* Enrico Moretti & Daniel Wilson, *State Incentives for Innovation, Star Scientists and Jobs: Evidence from Biotech*, NAT’L BUREAU OF ECON. RESEARCH, Working Paper No. 19294 (Aug. 2013), available at <http://www.nber.org/papers/w19294.pdf>, at 4–5 (finding that state R&D tax incentives and biotech subsidies for local firms were correlated with gains in employment in the non-traded sector, including retail, construction and real estate, suggesting that “by increasing employment in biotech, the incentives indirectly increase employment in local services, like construction and retail, whose demand reflect the strength of the local economy”).

88. MORETTI, *supra* note 40, at 61–62 (explaining why the “high-tech multiplier effect” is so much larger than that of other industries).

89. MORETTI, *supra* note 40, at 97–99.

90. *Id.* at 98.

91. *Id.* at 99; *see also* Enrico Moretti, *Human Capital Externalities in Cities*, in HANDBOOK OF REGIONAL AND URBAN ECONOMICS (J. Vernon Henderson & Jacques F. Thisse eds., 2004).

92. MORETTI, *supra* note 40, at 55–63, 97–101.

form, local amenities, and political attitudes.”⁹³ As Part V will discuss further, this is precisely why state and local governments are so eager to grow and attract *innovation* clusters in U.S. regions—not, say, clusters devoted to paper making.⁹⁴

D. WHERE ARE THE INNOVATION CLUSTERS TODAY?

The final question to ask is where precisely these lucky innovation clusters are located. Some places, like Silicon Valley and Boston’s Route 128, have become so strongly associated with innovation in the popular imagination that further inquiry seems unnecessary.⁹⁵ But how do we identify the others, and what metric do we use to distinguish an innovation cluster from any other region?

There are various ways to observe more precisely where the innovation clusters are today.⁹⁶ Patent counts and patents per capita is the most obvious indicator because they tell us where the innovators are probably located.⁹⁷ The five states that generate the most patents are California, New York, Texas, Washington, and Massachusetts.⁹⁸ The top four states (California, New York, Texas, and Washington) generate almost half of the patents granted in the United States.⁹⁹ A similarly striking disparity in patent counts is seen at the level of metropolitan regions. As of 2013, around 63% of U.S.

93. MORETTI, *supra* note 40, at 77.

94. See, e.g., Muro & Katz, *supra* note 41, at 4 (noting that federal and state and local policymakers are embracing the idea that “regional innovation clusters” have the power to re-shape an economy).

95. “[T]hroughout human history we have observed that creative activity has been concentrated in certain places and at certain times; consider Florence under the Medici, Paris in the 1920s, England during the Industrial Revolution, Silicon Valley and even Wall Street in more recent times. For every generation,” Feldman and Dieter Kogler write, “there is some location that captures the imagination as a locus of creative activity and energy.” Feldman & Kogler, *supra* note 55, at 384.

96. Various metrics are used to measure innovation, including IP ownership. See GREENHALGH & ROGERS, *supra* note 7, at 62–63. Other metrics include start-up activity, early-stage venture capital, and employment in high-tech services. *Id.* at 63.

97. MORETTI, *supra* note 40, at 82. As indicators of innovation, patents are imperfect, mainly because patents are both over and under inclusive. Many innovations are not patented, and any patented inventions are never transformed into successful innovations. See, e.g., *id.* at 83 n.1.

98. For patent counts by state for 2015, see PATENT COUNTS BY ORIGIN AND TYPE CALENDAR YEAR 2015, USPTO, http://www.uspto.gov/web/offices/ac/ido/oeip/taf/st_co_15.htm.

99. MORETTI, *supra* note 40, at 83.

patents were developed by people living in just twenty metropolitan areas, home to only 34% of the U.S. population.¹⁰⁰

Another way to observe where innovation occurs, or is likely to occur, is to observe where people with a college degree are located. Here too, there is significant geographic disparity. Metro areas with the highest share of workers with a college degree include Stamford CT, Washington, DC, Boston, MA, Madison, WI, San Jose, CA, Ann Arbor, MI, Raleigh-Durham, NC, and San Francisco-Oakland, CA. Regions with the lowest share of college-educated workers include Merced, CA, Yuma, AZ, Flint, MI, and Houma-Thibodaux, LA.¹⁰¹

Assuming a strong correlation between patents and/or education and value-adding innovation, then these regions should be expected to experience significantly more of the economic benefits of innovation than other regions, even within the same country. The broader point is that innovation, while it obviously benefits society at large, *does not* benefit all geographic communities equally. People living in the innovation clusters above are likely to experience far more of the benefits—higher salaries, better public services, and generally higher standards of living—than people living outside the clusters. The implications of this reality are obviously tremendous. The remainder of this Article discusses the implications for innovation policy, and particularly for public financing of innovation.

III. INTELLECTUAL PROPERTY AND INNOVATION FINANCE

The subject of this Article, again, is how the geographic distribution of the benefits of innovation impacts innovation policy. What is innovation policy? As noted in the Introduction, scholars identify various mechanisms

100. See Jonathan Rothwell et al., *Patenting Prosperity: Invention and Economic Performance in the United States and its Metropolitan Areas*, BROOKINGS INST. (Feb. 2013), at 12–13, <https://www.brookings.edu/research/patenting-prosperity-invention-and-economic-performance-in-the-united-states-and-its-metropolitan-areas/>. Top patent per capita metro regions include the San Jose-Sunnyvale-Santa Clara, CA (the leader by far), San Francisco-Oakland-Fremont, CA, New York-Northern New Jersey-Long Island, NY-NJ-PA, Los Angeles-Long Beach-Santa Ana, CA, Boston-Cambridge-Quincy, San Diego-Carlsbad-San Marcos, CA, and Seattle-Tacoma-Bellevue, WA. For patent counts at the level of metropolitan area for 2013, see PATENTING IN TECHNOLOGY CLASSES, BREAKOUT BY ORIGIN, U.S. METROPOLITAN AND MICROPOLITAN AREAS, USPTO, http://www.uspto.gov/web/offices/ac/ido/oeip/taf/cls_cbsa/allcbsa_gd.htm. See also MORETTI, *supra* note 40, at 83–85.

101. MORETTI, *supra* note 40, at 94–95.

through which government can influence innovation.¹⁰² This Article restricts itself to the two major forms of innovation incentive: innovation finance and intellectual property. This section lays out the purpose and operation of each form of incentive, and the most commonly discussed costs and benefits of each. Part IV then discusses how geography influences which level of government should be responsible for which type of incentive.

A. GOVERNMENT INCENTIVES FOR INNOVATION, GENERALLY

As explained in the last section, innovation benefits society in many ways. But the common wisdom is that, absent incentives, people and firms will under-invest in research, invention and commercialization primarily as a result of the difficulty of appropriating the full value of their productions.¹⁰³ The basic problem is that “an innovation can benefit more people and companies than just the innovating firm,” and the inability to internalize the full benefits of an innovation “may lead to an undersupply.”¹⁰⁴ There are two common theoretical frameworks used to conceptualize this market failure.¹⁰⁵ The first is that the knowledge generated through innovation is a “public good”: nonrival and nonexcludable and thus likely to be undersupplied by the private sector.¹⁰⁶ However, although this may be true for a mathematical theorem or basic research performed in a public institution with no near-term application, it is often inaccurate to say that knowledge is nonrival and nonexcludable, especially with respect to applied knowledge in a highly competitive setting where various exclusion mechanisms are possible.¹⁰⁷

102. See SCOTCHMER, *supra* note 6, at 242–43.

103. See, e.g., GREENHALGH & ROGERS, *supra* note 7, at 17–23 (explaining various market failure in the production of innovation, including public goods, externalities, imperfections in capital markets, and unproductive racing).

104. *Id.* at 17–18.

105. “Market failure” is used to describe a situation where there is a strong possibility that the market, guided only by the actions of private actors, will not lead to the optimal outcome. See GREENHALGH & ROGERS, *supra* note 7, at 18.

106. The basic definition of a public good is that it is nonrival (one person’s consumption of it does not interfere with another’s) and nonexcludable (others cannot efficiently be excluded from using the public good). MANKIW, *supra* note 37, at 218–21 (discussing typical kinds of public goods and explaining the “free-rider” problems associated with public goods). For IP scholarship using the public good justification see, for example, Wagner, *supra* note 32, at 1005–07.

107. Several sources note the ways in which knowledge in the commercial world does not behave as a public good at all. GREENHALGH & ROGERS, *supra* note 7, at 18–20. Once applied in a commercial product, knowledge can be quite rivalrous: imitation by a competitor, even though it does not deplete the value of the knowledge itself, depletes the

The second framework, which is superior for discussions of most commercial innovations, is that innovation creates positive externalities (or spillovers) for others that are not taken into account in private investment decisions.¹⁰⁸ The purpose of intellectual property regimes like patents is said to be “to correct this externality by more closely aligning the private and social value of producing new information.”¹⁰⁹ By giving inventors the chance to patent their inventions, patents increase incentives to invest in research and invention, and thereby indirectly increase their incentives to commercialize patented ideas.¹¹⁰

But intellectual property rights are not the only incentive government can use to spur investment in innovation. Governments also use innovation finance: public financing for innovation drawn from public revenues.¹¹¹ Innovation finance has the same general purpose as IP—to fund innovation in light of market failure and correct for the externalities associated with innovation—except rather than creating property rights, government uses public money to make up the difference between the social returns and the private returns from innovation.¹¹² Examples of innovation finance include research grants, prizes, tax credits, public venture capital, investments in education and working training, and other direct expenditures of public money on the innovation enterprise.¹¹³

profits of the first producer of the knowledge. *Id.* at 19. Many kinds of knowledge can be excluded even absent IP protection (e.g. through secrecy). *Id.* at 19–20; *see also* Tim Wu, *Properties of Information & the Legal Implications of Same*, COLUMBIA CTR. FOR L. AND ECON. STUDIES, Working Paper No. 482 (June 2014) (discussing recent literature noting the limits to information’s status as a public good).

108. GREENHALGH & ROGERS, *supra* note 7, at 20; *see also* Duffy, *supra* note 20, at 693 (noting that the patent system attempts to “account for the positive externalities associated with the creation of technical information”).

109. Duffy, *supra* note 20, at 694.

110. *See, e.g., id.* at 693–94.

111. *See, e.g.,* SCOTCHMER, *supra* note 6, at 242–43.

112. SCOTCHMER, *supra* note 6, at 242–43; GREENHALGH & ROGERS, *supra* note 7, at 24–25.

113. The literature on non-IP incentives to invest in innovation is vast and growing. *See, e.g.,* Frischmann, *supra* note 5, at 377–92 (discussing different forms of incentives including tax, research grants, and procurement); SCOTCHMER, *supra* note 6, at 40–46 (discussing innovation prizes); Hemel & Ouellette, *supra* note 5, at 303 (discussing prizes, grants, and R&D tax credits); Hrdy, *Commercialization Awards*, *supra* note 5, at 51–72 (discussing state and federal venture capital programs); *see also* JAMES BESSEN, *LEARNING BY DOING: THE REAL CONNECTION BETWEEN INNOVATION, WAGES, AND WEALTH* 19–20 (2015) (discussing the need for investment in education and other incentives for encouraging “broad based learning of new technical skills”).

B. MAJOR DIFFERENCES BETWEEN INTELLECTUAL PROPERTY AND INNOVATION FINANCE

The major distinctions between IP and innovation finance have been discussed by economists who study innovation such as Suzanne Scotchmer,¹¹⁴ and drawn out recently by IP scholars like Amy Kapczynski and Lisa Larrimore Ouellette.¹¹⁵ There are two dimensions to these discussions: efficiency and fairness.

1. *Efficiency*

From the perspective of efficiency, the fundamental difference between IP and innovation finance is that IP does not involve an expenditure of public funds; IP relies on innovators and private markets to determine the technical and commercial value of an innovation. As Scotchmer put it, the “lure of intellectual property” as opposed to “public sponsorship” of innovation is that IP automatically “tap[s] ideas for invention that are widely distributed among firms and inventors.”¹¹⁶ Government does not have to tell people what or whether to innovate; it just creates the possibility of obtaining an IP right for a qualifying innovation.¹¹⁷

On the other hand, IP rights create deadweight loss in the sense that some people cannot afford to pay for IP-protected goods who otherwise could.¹¹⁸ This deadweight loss is the “static,” short-term inefficiency of IP: where consumers cannot buy what they otherwise could in an unaltered market.¹¹⁹ IP also creates “dynamic,” longer term inefficiencies because consumers of the innovation may themselves be innovators, and the existence of IP rights may prevent them from researching or marketing cumulative innovations that otherwise would have benefited society.¹²⁰

114. See SCOTCHMER, *supra* note 6, at 37–40.

115. Kapczynski, *supra* note 5, at 970–80; Hemel & Ouellette, *supra* note 5, at 303; *see also* Frischmann, *supra* note 5, at 348–92.

116. SCOTCHMER, *supra* note 6, at 38.

117. *Id.*; *see also* Hemel & Ouellette, *supra* note 5, at 303 (observing that patents are “market-set”).

118. As Kapczynski explains, “Because information should ideally be priced at zero, any positive price generates static (short term) inefficiency, which economists refer to as deadweight loss. This kind of net loss of social welfare ‘occurs when people are excluded from using the good even though their willingnesses to pay are higher than the marginal cost.’” Kapczynski, *supra* note 5, at 982 (quoting SCOTCHMER, *supra* note 6, at 36).

119. *Id.*

120. *Id.* (“Positive price compromises not only static but also dynamic efficiency because information is an input and output of its own production process.”); *see also* Suzanne Scotchmer, *Standing on the Shoulders of Giants: Cumulative Research and the Patent Law*, 5 J. ECON. PERSPECTIVES 29, 32–35 (1991) (asserting that patent incentives for initial creators can impede cumulative innovation by second-generation creators).

Another problem, also highlighted in Kapczynski's work, is that IP rights, and patents in particular, can distort investment in innovation at a structural level by awarding some forms of innovation but not others or by awarding some forms of innovation more than others.¹²¹ For example, patents have a limited term length and so may be less valuable for innovations that take a long time to test and develop; patents cannot be obtained for publicly disclosed innovations¹²² and do not provide an incentive to innovate when the exclusive right is virtually impossible to enforce; and patents serve little purpose when there is no private market at all absent government procurement.¹²³ The risk of distortion of innovation, along with the usual concern about deadweight loss, has led scholars like Kapczynski to be skeptical that exclusive rights, on their own, will ensure efficient production of new information.¹²⁴

In contrast, innovation finance avoids the static and dynamic inefficiencies of creating exclusive rights; applies to a broader variety of innovations, including those that are best kept secret or otherwise hard to protect through IP; and addresses multiple market failures, including where innovators have trouble raising money to finance their operations.¹²⁵ Perhaps most importantly, because innovation finance is not dependent on

121. See Kapczynski & Syed, *The Continuum of Excludability and the Limits of Patents*, 122 YALE L.J. 1900, 1942 (2013) (concluding “that a patent system will predictably and systematically distort private investment decisions regarding innovation, overstating the value of highly excludable information goods and understating the value of highly nonexcludable ones,” and thus “fail to provide sufficient private returns to enable investment in certain information goods that clearly offer a net social benefit”). For a detailed discussion of the problem of patent-caused distortion of innovation in the pharmaceutical industry, see Sachs, *supra* note 8, at 8–19.

122. For example, some argue that drug companies lack sufficient incentives to innovate in manufacturing methods in part because they must rely mainly on secrecy to protect these methods. Nicholson Price, *Making Do in Making Drugs: Innovation Policy and Pharmaceutical Manufacturing*, 55 B.C. L. REV. 491, 523 (2014) (arguing that “the unique aspects of trade secrecy—including its practical limitations, an unbounded timeframe, process specificity, and limitations on personnel—make it structurally less capable of incentivizing pharmaceutical innovation”).

123. See Sachs, *supra* note 8, at 8–19 (identifying and discussing three forms of patent-caused distortion in the pharmaceutical industry); see also Camilla A. Hrdy, *Rachel Sachs: Prizing Insurance*, WRITTEN DESCRIPTION (May 31, 2016), <http://writtendescription.blogspot.com/2016/05/rachel-sachs-prizing-insurance.html>.

124. See Kapczynski & Syed, *supra* note 121, at 1960 (“If, as we have shown, property rights in information are themselves potentially distorting, then even if our sole aim is to achieve efficiency, we cannot assign decisions about allocation solely to the market.”).

125. For example, while patents address the appropriability problem, they do not directly address credit constraints in commercialization. Public financing for new companies may be superior to commercialization patents in this respect. See Hrdy, *Commercialization Awards*, *supra* note 5, at 43–51.

signals from the market, it can accomplish something IP cannot: promote innovations with broad social consequences that are not necessarily profitable enough to be attractive under an IP-only regime. In this regard, scholars have explored many options besides just grants and prizes. For example, Peter Lee explains in a recent article that IP rights are especially ineffective at promoting “social innovations” in areas like health, safety, education, and environmental protection. Instead, social innovations rely largely on non-market mechanisms such as government funding and charitable markets.¹²⁶

These strengths have to be weighed against the fact that innovation finance requires some level of government involvement to determine how much the innovation is worth and how much public money government should spend on its success. Daniel Hemel and Lisa Larrimore Ouellette thus refer to innovation finance mechanisms such as grants and prizes as “government-set” rather than “market-set” rewards for innovation: they rely on public officials rather than decentralized markets and the forces of supply and demand to figure out how much a particular technology should be subsidized.¹²⁷ When the government has more information than the private sector, then government has an easier time making this determination.¹²⁸ But when the government does not have more information than the private sector, this necessarily requires public officials at some level “to figure out how much a particular technology should be subsidized” without the guidepost of supply and demand.¹²⁹ Commentators have proposed various ways that government could link government rewards with market returns. For instance, before funding research, government could require innovators to obtain matching funds from private investors to mitigate the risks associated with choosing between different projects or different avenues or research.¹³⁰ But the mechanisms are by no means perfect. Thus, just like

126. See Lee, *supra* note 5, at 1–10. More recently, Rachel Sachs has argued that prescription drug insurance could be turned into a non-market mechanism for promoting pharmaceutical innovation. See Sachs, *supra* note 8, at 34–41 (arguing that prescription drug insurance could be wielded as an innovation incentive in a way that addresses the “market-shaped innovation distortions” of patent law).

127. Hemel & Ouellette, *supra* note 5, at 327. Hemel and Ouellette’s insight is that some forms of innovation finance, such as R&D tax credits, are more dependent on market signals than others.

128. Brian D. Wright, *The Economics of Invention Incentives: Patents, Prizes, and Research Contracts*, 83 AM. ECON. REV. 691, 691–92 (1983).

129. Hemel & Ouellette, *supra* note 5, at 327.

130. See Hrdy, *Commercialization Awards*, *supra* note 5, at 14, 58–59 (“In order to mitigate the risks associated with government ‘picking winners,’ awards require obtaining private sector matching before money changes hands.”).

with IP, innovation finance is likely to result in some deadweight loss, including wasted public money on bad investments and distortion of innovation decisions in the private sector.¹³¹

2. *Fairness*

According to Scotchmer, IP is also likely to be fairer in many cases because innovations that are covered by IP rights are ultimately financed directly by users of the innovation rather than general taxpayers.¹³² As she put it, with IP, “[e]ach innovation is paid for voluntarily through proprietary prices.”¹³³ In contrast, innovation finance mechanisms like procurement and prizes will often draw on general taxpayer revenues to finance innovations that benefit some taxpayers but not others. While public funding for technologies that help entire populations, such as vaccines for common diseases, may be uncontroversial, in many situations, “different constituencies” may “argue for different R&D projects.”¹³⁴ For instance, many would probably prefer users of discretionary consumer products like music players to subsidize the research that supplies those products, rather than everyone who pays taxes.¹³⁵ As Scotchmer put it, “[t]axpayers might rightfully revolt if asked to bear the costs of developing, say, computer games.”¹³⁶ In other words, with IP, people agree to pay more for what they want rather than being forced to pay for what they do not want through the tax system.

However, IP rights create other fairness issues. As Kapczynski has pointed out, IP rights can challenge the principles of distributive justice, which is concerned with the allocation of resources across society.¹³⁷ Because IP rights tend to force consumers to pay more for products than they otherwise would in a naked market, the gains from IP-protected goods

131. Deadweight loss in this context is the fall in total welfare due to people allocating resources according to the incentive rather than the true costs and benefits of the innovation. MANKIW, *supra* note 37, at 159, 242; *see also* Kapczynski, *supra* note 5, at 980–81.

132. SCOTCHMER, *supra* note 6, at 38–39. For the argument that the “user pays” feature of IP comports with distributive justice, see Jeremy Sheff, “Who Should Pay for Progress? – IPSC Draft Talk” (Aug. 6, 2014), <http://jeremysheff.com/2014/08/06/who-should-pay-for-progress-ipsc-talk-draft/>.

133. SCOTCHMER, *supra* note 6, at 38; *see also* Hemel & Ouellette, *supra* note 5, at 346.

134. SCOTCHMER, *supra* note 6, at 39.

135. *See id.* at 38.

136. *See* Nancy Gallini & Suzanne Scotchmer, *Intellectual Property: When Is It the Best Incentive System?*, 2 INNOVATION POL’Y & ECON. 51, 55 (2002).

137. Kapczynski, *supra* note 5, at 995.

such as expensive pharmaceutical drugs are often unevenly distributed across society, limited to those people who can pay for access.¹³⁸ In theory, innovation finance can be used to mitigate this problem or to deliberately redistribute resources from one segment of the population to another. For example, if government directly finances research into diseases that are more frequently found in poorer populations, this could constitute a redistribution of resources from the wealthy to the poor.¹³⁹ On the other hand, some argue that IP rights themselves have a redistributive component. For example, Madhavi Sunder has argued that the expansion of global IP protections can serve as “a tool for protecting poor people’s knowledge” in undeveloped countries even as it protects the “knowledge and economic interests of the developed world.”¹⁴⁰

IV. THE ROLE OF JURISDICTION IN INNOVATION POLICY CHOICES

The upshot of the last section is that there are two ways to promote innovation and yet no principled way to choose between them. Both have costs and benefits, and in some cases they perform different functions. A previously unobserved dimension in this debate is the role of jurisdictional allocation in innovation policy choices. If optimizing efficiency in innovation policy is a goal, a good place to start is to consider which jurisdiction should be responsible for innovation policy and, more specifically, for *which aspects* of innovation policy. This section argues that, within the United States, which jurisdiction is responsible—national or subnational—depends in part on which policy tool is selected—intellectual property rights or innovation finance. In turn, which policy mechanism we choose may influence which jurisdiction is responsible. This section will show that if patents are selected, jurisdiction is more likely to be national or even global. If innovation finance is selected, however, jurisdiction should often be local.

138. *See id.* at 993–96.

139. *See* MADHAVI SUNDER, FROM GOODS TO THE GOOD LIFE 174–78 (2012) (discussing the downsides of patents in supplying life-saving drugs to the poor, including pricing out poor consumers and insufficient research into “neglected diseases” that mainly afflict poor populations); *see also* Kapczynski, *supra* note 5, at 1001–02 (arguing that government procurement including prizes and contracts are promising alternatives to IP in distributive terms and might be used to facilitate access for the poor).

140. *See, e.g.*, SUNDER, *supra* note 139, at 126–44 (arguing that IP rights such as patents, copyrights, and geographic indicators can provide a tool for protecting the knowledge and cultural heritage of poor people in less developed economies).

A. COLLECTIVE ACTION FEDERALISM

In addressing the question of jurisdictional allocation, this section relies on the collective action theory of federalism, which appears in Cooter and Siegel's oft-cited article, *Collective Action Federalism: A General Theory of Article I, Section 8*.¹⁴¹ The basic rule is that the scope of national authority depends on whether there is a "collective action failure" that prevents states from resolving the problem in an efficient manner on their own. Whether a collective action failure exists depends on whether state action would produce severe inter-jurisdictional externalities (spillovers) that prevent effective resolution of the public problem and that states cannot resolve through inter-state negotiations because the transaction costs are very high.¹⁴² Otherwise, when no collective action failure exists, we should follow the "internalization principle": assign power to the smallest subnational government that internalizes the benefits of its policies without creating spillovers for other regions.¹⁴³

This theory is not new but is derived from longstanding literature on economic federalism.¹⁴⁴ Cooter and Siegel's innovation is to simplify the framework into two basic concepts—collective action failure and the internalization principle—and to argue that these concepts explain and justify the basic structural framework of the U.S. Constitution, in which the states possess all powers they had prior to ratification of the Constitution unless those powers are divested, expressly or implicitly, by the national powers specifically enumerated in Article I, Section 8.¹⁴⁵ So while this

141. Cooter & Siegel, *supra* note 14, at 118–119; *see also* Neil Siegel, *Collective Action Federalism and its Discontents*, 91 TEX. L. REV. 1937 (2013).

142. Cooter & Siegel, *supra* note 14, at 118–19, 139–40 (explaining the relevance of transaction costs for their theory).

143. *Id.* at 137.

144. For examples of scholarship discussing common economic theories of federalism (often called simply "economic federalism"), *see* Robert P. Inman & Daniel L. Rubinfeld, *Rethinking Federalism*, 11 J. ECON. PERPS. 43, 45 (1997) ("The principle of economic federalism prefers the most decentralized structure of government capable of internalizing all economic externalities, subject to the constitutional constraint that all central government policies be decided by an elected or appointed 'central planner.'"); Wallace E. Oates, *An Essay on Fiscal Federalism*, 37 J. ECON. LIT. 1120, 1122 (1999) (stating that the "basic principle of fiscal decentralization" is "the presumption that the provision of public services should be located at the lowest level of government encompassing, in a spatial sense, the relevant benefits and costs"). *See also* Richard A. Posner, *Toward an Economic Theory of Federal Jurisdiction*, 6 HARV. J.L. & PUB. POL'Y 41, 41–50 (1982) (applying economic federalism to address the optimal allocation of authority between federal and state courts).

145. *See* Cooter & Siegel, *supra* note 14, at 118–119; U.S. CONST. art. I, § 8, amend. X. Jack Balkin provides a similar justification for the national government's authority to

Article uses Cooter and Siegel's terminology, it actually relies on the body of economic federalism theory itself.

Economic federalism theory typically seeks to determine the optimal, most efficiency-enhancing allocation of authority between federal and state governments with respect to substantive areas of law.¹⁴⁶ But it has also been adopted, for example, to decide whether jurisdiction should be assigned to federal or state courts.¹⁴⁷ In that context, Richard Posner explains economic federalism in very similar terms to Cooter and Siegel, writing that the economic theory of federalism contains “a presumption in favor of shifting governmental power from higher to lower levels, from broader to narrower jurisdictions—for present purposes, from the federal to the state level.”¹⁴⁸ (This is the internalization principle prescribing local jurisdiction.) However, “If either the benefits or the costs of a governmental action are felt outside the jurisdiction where the action is taken, and the costs of negotiations between governments are assumed . . . to be very high, then there is a strong argument that the responsibility for the action should be assigned to a higher level of government with a broader jurisdiction.”¹⁴⁹ (This is the collection action failure necessitating federal jurisdiction.)

The economic federalism framework also closely resembles the “principle of subsidiarity” underlying the relationship between the European Union and its member states. “Under the principle of subsidiarity,” the European Union Treaty states, “in areas which do not fall within its exclusive competence, the Union shall act only if and in so far as the objectives of the proposed action cannot be sufficiently achieved by the Member States, either at central level or at regional and local level, but can rather, by reason of the scale or effects of the proposed action, be better achieved at Union level.”¹⁵⁰

The economic federalism framework has often been applied specifically to the financing of public goods. Here, economic federalism prescribes that the national government should be responsible for “national public goods”

regulate commerce under Article I, Section 8, Cl. 3. *See* JACK M. BALKIN, *LIVING ORIGINALISM* 140 (2011) (“Properly understood, the commerce power authorizes Congress to regulate problems or activities that produce spillover effects between states or generate collective action problems that concern more than one state.”).

146. *See, e.g.*, Inman & Rubinfeld, *supra* note 144; Oates, *supra* note 144.

147. Posner, *supra* note 144, at 41 (“The specific question I want to address today is how economics, and specifically the economic theory of federalism, can aid decisions with respect to [the] allocation [of responsibilities between state and federal courts.]”).

148. *Id.* at 45.

149. *Id.* at 45.

150. Consolidated Treaty on European Union art. 5, Dec. 13, 2007, 2012 O.J. (C 326) 1. Thank you to Jean Galbraith and Lee Anne Fennell for this point.

that implicate inter-jurisdictional spillovers, such as national security.¹⁵¹ In contrast, state and local governments should be responsible for “local public goods” that are associated with a physical location and mainly affect a local population. Typical examples of local public goods include parks and fire stations.¹⁵² Unlike national public goods, which are said to be “pure” public goods because they are characterized by a high degree of nonrivalry and nonexcludability, local public goods are not necessarily nonexcludable or even nonrival. At some point, local public goods may suffer from congestion (crowding), meaning that one person’s enjoyment of the good detracts from others’ enjoyment.¹⁵³

B. APPLICATION TO INNOVATION POLICY

We can now apply economic federalism theory to determine which level of government should be responsible for which aspect of innovation policy. The simple question to ask is: what is the smallest level of government that internalizes the benefits of an IP regime, on the one hand, and innovation finance on the other?

When it comes to IP rights like patents, national jurisdiction is generally required.¹⁵⁴ In most cases, IP creates too many spillovers—cross-border copying of inventions and creative productions—for states to protect on their own. Subnational governments like states have a difficult time appropriating the benefits of new knowledge generated in innovation through patent rights. The reason is that patents, like real property rights, operate through the mechanism of exclusion, and are only valuable to the extent that they *actually prevent others from using the protected subject matter in the relevant market*.¹⁵⁵ Once the actual or potential market for an

151. See COOTER, *supra* note 14, at 105–07; Cooter & Siegel, *supra* note 14, at 137–38. For a fuller explanation of the difference between national and local public goods and different mechanisms for supplying them, see DENNIS MUELLER, CONSTITUTIONAL DEMOCRACY 81–83 (1996) (discussing public goods at the local, city, regional, national, and global level); JOSEPH STIGLITZ, ECONOMICS OF THE PUBLIC SECTOR 733–34 (3d ed. 2000) (using examples to illustrate the difference between local public goods and national public goods).

152. See COOTER, *supra* note 14, at 106; Cooter & Siegel, *supra* note 14, at 137.

153. See COOTER, *supra* note 14, at 105–07; Cooter & Siegel, *supra* note 14, at 137–38.

154. See Cooter & Siegel, *supra* note 16, at 149 (“Because the problem of unauthorized use extends across state lines, the problem is national and Congress is better placed than the states to solve it.”).

155. As noted above, IP scholars disagree over the *degree* to which innovators must internalize the benefits of their innovations, but, in the common utilitarian framework, the accepted role of IP is to assist creators in internalizing the benefits of their innovations by giving them reasonably effective rights of exclusion. Compare, e.g., Wagner, *supra* note 32, at 1033 (arguing that strong intellectual property rights in information “expands the

invention become interstate, state patent laws with limited jurisdictional reach simply cannot perform this function without risking out-of-state copying of their inventors' patented inventions.¹⁵⁶ Thus, if we choose patents as the innovation incentive, we are also necessarily choosing national, and probably eventually global, jurisdiction.¹⁵⁷

However, local governments *can* appropriate at least some of the benefits of innovation using innovation finance. Innovation finance incentives, such as research grants, prizes, and tax credits, do not operate through the mechanism of exclusion. Thus, the threat of instantaneous cross-border competition does not make them inherently ineffectual.¹⁵⁸ Moreover, as discussed in detail in Part II, the immediate economic impacts of innovation tend to be geographically localized and concentrated in certain parts of the country. Although innovation produces various national benefits—including consumer surplus, long-term improvements in economic well-being, and knowledge spillovers with respect to information that is easy to transfer and that has widespread relevance in interstate markets—innovation also produces significant “internalities” for the jurisdictions in which innovation occurs, including high-paying employment, increased business activity, and increased tax revenues.¹⁵⁹ In theory, local governments can, through the mechanism of public finance, capture enough of these benefits to give them an incentive to finance innovation.¹⁶⁰

total sum of open information available for further technological, cultural, and social development”), with Frischmann & Lemley, *supra* note 32, at 258 (asserting that *complete* internalization of externalities is not necessary to optimize IP-based incentives to innovate).

156. See, e.g., Hrды, *State Patents as a Solution*, *supra* note 20, at 111 (observing that states cannot internalize all the benefits of the research and information produced by state patents); see also Hrды, *State Patent Laws in the Age of Laissez Faire*, *supra* note 25, at 67–70 (on the end of U.S. state patent laws in the wake of interstate commerce).

157. As John Duffy has observed, externalities—the risk of cross-border copying and the flow of knowledge from one nation to another—is the primary justification for globally harmonized patent systems as well. See Duffy, *supra* note 20, at 693–700, 707–09.

158. Take a simple example. A state patent for an invention sold in interstate commerce is valueless or nearly valueless, assuming out-of-state competitors can quickly copy the invention and freely reproduce it. But the value of a state subsidy for the same invention, such as a prize or a research grant, is not valueless. It can provide an incentive to research the invention *ex ante* or compensate the inventor for her labor after the fact.

159. This Article borrows the term “internality” from Professor Cooter. See COOTER, *supra* note 14, at 109; see also *supra* Section II.C.

160. As in the case of IP rights, one can argue over the degree to which states must internalize the benefits to adopt policies that support local innovation. This issue is discussed further *infra* Section IV.C.3.

Therefore, if one takes seriously the internalization principle and the economic federalism literature from which it is derived, state and local governments should sometimes be responsible for financing innovation as a local public good. If one chooses innovation finance as the innovation incentive, one may also necessarily be choosing a local jurisdiction.

C. THE BENEFITS OF LOCAL JURISDICTION IN INNOVATION FINANCE

There are various advantages to choosing local rather than national jurisdiction in innovation finance. Proponents of economic federalism give various efficiency-based justifications for adopting a baseline preference for localism in the absence of significant externalities. Richard Schragger identifies three common justifications: local governments' stronger incentives to pursue policies that provide for a healthy local economy; the efficiency-promoting effects of inter-jurisdictional competition for mobile residents; and the greater propensity of state and local governments to take risks and adopt "policy experiments."¹⁶¹ Without taking a stand on whether these views are correct, this Article briefly explains each below and how it may apply in the context of innovation finance.

1. *Better Incentives to Design Effective Local Innovation Policies*

The main efficiency-based justification that Cooter and Siegel provide for preferring a local jurisdiction in the absence of severe spillovers is that local officials are likely to have better incentives to supply, and design policies relating to, local public goods in a manner that meets residents' needs and preferences.¹⁶² Specifically, Cooter and Siegel argue that residents "possess better information than nonresidents" when it comes to situating and scaling local public goods, and "stronger incentives than nonresidents to monitor the officials responsible for creating and maintaining" those goods; in turn, local officials have "better incentives than central officials" to supply the goods in response to residents' demands.¹⁶³

161. See Richard C. Schragger, *Decentralization and Development*, 96 VA. L. REV. 1837, 1853–63 (2010); see also, e.g., Jonathan Adler, *Interstate Competition and the Race to the Top*, 35 HARV. J.L. & PUB. POL'Y 91–92 (2013) (discussing the economic benefits of inter-jurisdictional competition); John C. Yoo, *The Judicial Safeguards of Federalism*, 70 S. CALIF. L. REV. 1311, 1402–03 (1997) (citing literature on economic federalism to justify preserving federalism).

162. Cooter & Siegel, *supra* note 14, at 137–38 (explaining the claim that local officials should have better incentives than central officials with respect to supplying many local public goods); see also COOTER, *supra* note 14, at 106.

163. Cooter & Siegel, *supra* note 14, at 138; see also COOTER, *supra* note 14, at 106.

An unstated assumption of Cooter and Siegel's theory is that local officials must seek to satisfy the needs of local residents because otherwise residents will object and/or leave the jurisdiction.¹⁶⁴ On this view, local governments are motivated largely by the threat of flight to other U.S. (or foreign) jurisdictions. In this sense, Cooter and Siegel's theory resembles a slightly distinct theory of federalism, called "market-preserving federalism," popularized by Barry Weingast.¹⁶⁵ On this view, "inter-jurisdictional competition provides political officials with strong fiscal incentives to pursue policies that provide for a healthy local economy."¹⁶⁶ More specifically, Weingast explains, "As long as capital and labor are mobile, market-preserving federalism constrains the lower units in their attempts to place political limits on economic activity, because resources will move to other jurisdictions."¹⁶⁷

The main implication of these views for innovation policy are that subnational governments should have exceptionally strong incentives to take actions that foster innovation at the regional level to the extent that the economic impacts of innovation are concentrated in the region and do not spill over significantly to other states.¹⁶⁸ Local governments certainly attempt to foster regional innovation economies in practice. States typically have economic development offices whose primary responsibility is to design innovation incentives that will promote economic development, and

164. Note that the "threat of flight" is not necessary for Cooter and Siegel's view to hold. As Schragger notes, "The threat of exit need not be the only mechanism [motivating governments]; presumably electoral pressure will induce officials to act one way or another." Schragger, *supra* note 161, at 1854. That said, the notion that local officials are motivated by the threat of flight to other U.S. jurisdictions certainly strengthens Cooter and Siegel's argument that local officials have stronger incentives than the national government to satisfy resident voters' demands. See Cooter & Siegel, *supra* note 14, at 138 ("[L]ocal officials have better incentives than central officials for supplying local public goods.").

165. See, e.g., Barry Weingast, *Second Generation Fiscal Federalism*, 65 J. URB. ECON. 279, 280–82 (2009); see also Schragger, *supra* note 161, at 1853–54 (explaining the core tenets and assumptions of market-preserving federalism).

166. See Schragger, *supra* note 161, at 1853 (quoting Weingast, *supra* note 165, at 281).

167. Barry Weingast, *The Economic Role of Political Institutions: Market-Preserving Federalism and Economic Development*, 11 J. L. ECON. & ORG. 1, 5 (1995). As Schragger notes, this particular view—that "the cost of capital flight drives officials' behavior"—relies on the assumption that the national government does not experience the same level of inter-jurisdictional competition, presumably because the threat that residents will leave the United States altogether is not as realistic. Schragger, *supra* note 161, at 1853. *But see id.* at 1854–55 (questioning whether local officials really have stronger incentives to provide for a healthy local economy than central government officials).

168. See Cooter & Siegel, *supra* note 14, at 138; see also COOTER, *supra* note 14, at 106.

to collect information and keep reports about ongoing progress and challenges to economic development in the region.¹⁶⁹ As Part V will discuss, the pervasiveness and variety of local innovation incentives is staggering.

A further implication of Weingast's view that the threat of flight motivates subnational governments is that local innovation policies should be more sensitive than federal innovation policy to the relatively near-term demands of the mobile actors that make up an innovation economy. They should be specifically designed to attract venture capital, research corporations, innovative start-ups, university faculty, highly skilled talent, and all the other mobile ingredients of a cluster.¹⁷⁰ As already alluded to, this holds true in practice. The explicit goal of U.S. state innovation incentives is to spur jobs and business activity *in the region*, potentially at the expense of others. As the next section will discuss, this is quite different from the tenor of federal innovation policy, which is more focused on the production of national public goods (mainly national security) and the production of information whose near-term economic value is not always clear.¹⁷¹

2. *Tiebout Clustering*

A slightly different view is that local jurisdiction should be preferred in the absence of severe spillovers because inter-jurisdictional competition leads to more efficient provisioning of local public goods and efficient sorting of residents into appropriate locations.¹⁷² In an influential article, Charles Tiebout, hypothesized that local governments supply local public goods based on the demands of mobile "consumer voters," who can "shop" among local jurisdictions for the community that best satisfies their preferences; therefore local jurisdictions will provide public goods that meet those preferences as efficiently as possible.¹⁷³ The thrust of this argument, as one economist puts it, is that "[c]ompetition among communities . . . will result in communities' supplying the goods and

169. See, e.g., MICHIGAN ECON. DEV. CORP., *Transparency*, <http://www.michiganbusiness.org/about-medc/transparency/> (last visited May 4, 2016); TEXAS ECON. DEV. CORP., *Reports, Directories & Databases*, <https://texaswideopenforbusiness.com/re/sources/reports-directories> (last visited, May 4, 2016).

170. See Weingast, *supra*, note 165, at 281.

171. See discussion *infra* Section V.B.1.

172. See Inman & Rubinfeld, *supra* note 144, at 46 (explicating Tiebout's hypothesis); STIGLITZ, *supra* note 151, at 734–36 (also explicating Tiebout's hypothesis).

173. Charles M. Tiebout, *A Pure Theory of Local Expenditures*, 64 J. POL. ECON. 416, 418–20 (1956).

services individuals want and producing these goods in an efficient manner.”¹⁷⁴

According to Cooter, the major benefit of the Tiebout hypothesis is that “sorting diverse populations into groups with relatively homogenous tastes can give each of [those groups] their preferred public goods.”¹⁷⁵ Cooter refers to this as “clustering”—where “[p]eople with similar tastes voluntarily cluster together in order to enjoy their preferred combination of local public goods.” For example, if one person values parks more than safe streets, and another values safe streets more than parks, each person can move to the locality that specializes in parks or safe streets, respectively.¹⁷⁶

Importantly, the Tiebout hypothesis is subject to major limitations, including that people are perfectly mobile and have full knowledge of the differences between possible jurisdictions; and many doubt whether efficient sorting of mobile residents occurs in practice.¹⁷⁷ But nonetheless the Tiebout hypothesis has been used to support the assumption that a “competitive, decentralized system of local government goods provision would be more tailored to local conditions than a centralized one,” and that this configuration contributes to economic growth in regions such as the United States with federal constitutional systems and local government autonomy.¹⁷⁸

Tiebout clustering has obvious application in the context of innovation clusters. According to the theory, mobile residents—venture capital, research corporations, innovative start-ups, university faculty, skilled talent—will sort themselves into innovation clusters based on their shared preferences, and subnational governments will then tailor their programs to meet those preferences in competition with other states. Tiebout’s competitive location market should be more favorable to the creation of innovation clusters than a purely centralized policy.¹⁷⁹

174. STIGLITZ, *supra* note 151, at 735.

175. COOTER, *supra* note 14, at 129.

176. *Id.*

177. Tiebout made clear that his theory only applied if certain assumptions are made—most importantly, that “consumer-voters are fully mobile and will move to that community where their preference patterns, which are set, are best satisfied.” Tiebout, *supra* note 173, at 419; *see also* Schragger, *supra* note 161, at 1857–59 (discussing the major limitations and critiques of the Tiebout hypothesis).

178. Schragger, *supra* note 161, at 1858 (describing and critiquing this assumption); *id.* at 1837 (challenging the thesis that political decentralization in America has promoted economic development).

179. I am hardly the first to make this argument. Maryann P. Feldman, whose empirical work Section V.B.2 discusses, has been making this argument for years. *See, e.g.*, Maryann P. Feldman, *Location, Location, Location: Creating Innovation Clusters*, DEMOCRACY: A

Take the following example. All subnational governments that are interested in creating a certain type of technology cluster—say, electric vehicles—can compete with one another to design policies they believe will work best, crafting tailored incentives and supplying appropriate infrastructure. Meanwhile, other states can specialize in nurturing other technology areas where they see a comparative advantage: solar in Arizona, aviation in Illinois, medical research in Connecticut, and so on. Firms and talent working in these technology areas (Tiebout’s “consumer-voters”) can “vote with their feet,” locating in the places whose goals and policies suit their particular needs. The result would—theoretically—be higher-performing innovation clusters, and more innovation overall.

While it is not *necessary* to believe in the Tiebout hypothesis to support local jurisdiction over innovation policy—because, as mentioned, there are thought to be other benefits from relying on local officials to craft local policy—the Tiebout hypothesis does provide another justification for preferring local government when possible, and may have particular relevance if, as cluster policy advocates contend, innovation really does proceed more efficiently in clusters.¹⁸⁰

Importantly, as Schragger observes, efficient clustering via the Tiebout model can be obtained without full political decentralization of innovation policy.¹⁸¹ Rather than requiring each subnational jurisdiction to fund and independently design its own innovation policies, the federal government could provide funding to local governments on a competitive basis and provide guidelines for them to follow in crafting their policies. For example, former Michigan Governor Jennifer Granholm has spoken in favor of a new “Race to the Top” program in which states would compete for federal funds to incubate clean energy innovation in sectors like solar, wind, or electric vehicles.¹⁸² Winners would use the funds to create various clean energy

JOURNAL OF IDEAS (No. 21, Summer 2011) (“One of the strengths of American federalism is that each level of government has a role to play. In the case of clusters, regional and local agencies are better able than federal entities to tailor programs to the specific needs of local industry.”).

180. See discussion of cluster theory in Section II.C.

181. See Schragger, *supra* note 161, at 1858. That said, Cooter and Siegel stress the importance of fiscal fairness—the notion that local public goods should be financed using local taxes so that the beneficiaries of local public goods pay for them. Cooter & Siegel, *supra* note 14, at 138 (“[A] local public good can be financed by a local tax, which primarily hits the beneficiaries and misses nonbeneficiaries.”). This coincides with the “benefits principle” of taxation: the idea that people should pay taxes based on the benefits they receive from government services. See MANKIW, *supra* note 37, at 246.

182. See Anthony Flint, *Could We Model a National Energy Policy on ‘Race to the Top’?*, THE ATLANTIC: CITY LABS (Feb. 28, 2013), <http://www.theatlanticcities.com/>

innovation incentives, from tax credits to partnerships with state universities.¹⁸³ This competition could be expanded to cover the whole range of technology areas that the federal government thinks are important. As Granholm puts it, “Every state would have something to contribute.” California could focus on solar and Midwestern states on wind power, while Michigan could focus on electric car batteries.¹⁸⁴ I discuss this more cooperative approach to innovation policy in a separate article.¹⁸⁵

3. *Innovation Policy Experiments*

There is “another valid, and even more familiar, argument for preferring state government—the ‘experiments in separate laboratories’ argument.”¹⁸⁶ “Decentralization,” Richard Posner writes, is said to provide “valuable information about the provision of public services because diverse polities naturally hit on different solutions to common problems and the results of these different solutions can be compared.”¹⁸⁷

Several scholars, including myself, have pointed out the importance of experimentation in patent law and innovation policy in particular.¹⁸⁸ Lisa Larrimore Ouellette, for instance, has stressed the importance, in the patent law context, of promoting “innovation in promoting innovation.”¹⁸⁹ However, this kind of experimentation in innovation policy design will only realistically occur if jurisdictions are *first* able to reap the benefits of their policies. Why would regions bother to innovate if they cannot reap the

politics/2013/02/could-we-model-national-energy-policy-race-top/4829/ (quoting Granholm’s TED talk in Long Beach, CA).

183. See Jennifer M. Granholm, *A Jobs Race to the Top*, GOLDMAN SCHOOL OF PUB. POL’Y (Oct. 1, 2013), <http://gspp.berkeley.edu/news/news-center/a-jobs-race-to-the-top>.

184. Flint, *supra* note 182.

185. See Camilla A. Hrdy, *Cluster Competition*, 20 LEWIS & CLARK L.J. 981 (2016).

186. Posner, *supra* note 144, at 44; see also Schragger, *supra* note 161, at 1860 (“That state and local governments are valuable as laboratories of experimentation is a popular assumption . . .”) (citing *New State Ice Co. v. Liebmann*, 285 U.S. 262, 310–11 (1932)).

187. Posner, *supra* note 144, at 45.

188. See, e.g., Camilla A. Hrdy, *Dissenting State Patent Regimes*, 3 IP THEORY 78, 87 (2013) (arguing that one benefit of state patent regimes is that states could “generate troves of valuable information about the effects of patents in the marketplace and . . . begin to experiment with designing patent laws that work more effectively”); Hrdy, *State Patent Laws in the Age of Laissez Faire*, *supra* note 25, at 489 (“Especially given the continued move towards global uniformity, patent law could benefit from the policy experiments that divergent state patent regimes would produce, turning the states into decentralized ‘laboratories’ for improving the functioning of patent law.”); Lisa Larrimore Ouellette, *Patent Experimentalism*, 101 VA. L. REV. 65, 68 (2015) (arguing that U.S. patent policy “should focus not on uniformity, but rather on improving innovation incentives through an evidence-based approach that depends on policy diversity”).

189. Ouellette, *supra* note 188, at 68.

benefits of their policy experiments?¹⁹⁰ Consequently, like Cooter and Siegel, this Article focuses primarily on the issue of which government has the better incentives to adopt which kinds of policies—which, in turn, boils down to the question of which level of government is the smallest jurisdiction that can actually internalize the benefits of those policies.¹⁹¹

That said, assuming local governments do have sufficient incentives to experiment, this is another boon of localized governance in innovation policy. The diversity of programs discussed in Part V indicates that state governments in the United States *are* engaging in significant experimentation in terms of the amount, type, and particular design of innovation incentives. Moreover, there appears to be a significant amount of replication between states. At a National Academy of Sciences (NAS) symposium in 2009, local officials from a number of states (New York, Pennsylvania, Virginia, South Carolina, Kansas, Ohio, Washington, California, and Arizona) came together in Washington, D.C., to share specific strategies for growing innovation clusters. As described by rapporteur Charles Wessner in his report following the NAS symposium, “These initiatives can be seen as an ongoing experiments that can yield valuable insights on the role and limits of public policy in encouraging cluster-based economic growth and development.”¹⁹²

Maryann P. Feldman, whose work on “state science policy experiments” Section V.B.2(b) discusses, has explicitly stressed states’ propensity to engage in experimentation in innovation policy design and to learn from one another’s experiences. On the other hand, Feldman observes, replication is not always a boon. States’ tendency “to attempt to follow the same strategies or develop industries that are similar to those established in other places” has downsides.¹⁹³ “While it appears to be easier to follow the lead of another place,” Feldman writes,

190. Hrdy, *Dissenting State Patent Regimes*, *supra* note 188, at 83 (noting that state patents would generate “ ‘innovation spillovers’ for innovators and competitors across the country.”); *see also* Ouellette, *supra* note 188, at 86–87 (noting that “unconstrained ‘laboratories’ may under-innovate due to the externalities of both innovation itself (jurisdictions also do not internalize [all] the benefits of innovation policy) and innovation about innovation (jurisdictions also do not internalize the benefits of policy experiments)”); Brian Galle & Joseph Leahy, *Laboratories of Democracy? Policy Innovation in Decentralized Governments*, 58 EMORY L.J. 1333, 1335 (2009) (“State and local governments can be thought of as inventors without patents: because anyone can steal their new ideas, what incentive have they ever had to invent?”).

191. *See* Cooter & Siegel, *supra* note 14, at 137–38.

192. WESSNER, *supra* note 2, at 9.

193. Feldman, *Location, Location, Location*, *supra* note 179.

[T]his strategy certainly does not ensure success. Many emerging clusters within the United States and around the world often look to the successes of Silicon Valley and Route 128 as they attempt to promote their emerging economy. What can often be their downfall, however, is that these emerging clusters try to replicate the actions of the leading economies rather than fill a new niche by diversifying.¹⁹⁴

Therefore, one should take the benefit of more local innovation policy experiments with a grain of salt. As Wessner puts it in his remarks at the NAS symposium, it is helpful for local officials to learn what worked or failed elsewhere, and to extract “broader principles” from “analysis of the creation of clusters.”¹⁹⁵ Wessner identifies three such principles: the presence of localized knowledge, a skilled workforce, and the availability of capital; opportunities for entrepreneurship and collaboration; and the presence of “appropriate incentives.”¹⁹⁶ But, Wessner stresses, just as no two regions are identical, there is “no ready formula for recreating an innovation cluster.”¹⁹⁷

D. LIMITATIONS

The economic federalism model has significant limitations. The most obvious limitation is spillovers. As explained, the main justification for intervention by a national government is “collective action failure” among local governments caused by inter-jurisdictional externalities.¹⁹⁸ Even when a public good is seemingly as geographically localized as a bridge or a cobblestone street, “local governments *never* fully internalize the costs and benefits of their local economic policies.”¹⁹⁹ As shown below, local innovation policies can lead to externalities, both positive and negative, that necessitate national intervention through subsidy, preemption, or other regulations.²⁰⁰

194. *Id.*

195. WESSNER, *supra* note 2, at 8.

196. *Id.* at 5 (categorizing these ingredients as capabilities, opportunities, and incentives) (citing MAGGIONI, *supra* note 54).

197. *Id.*

198. See Cooter & Siegel, *supra* note 16, at 118–19; Inman & Rubinfeld, *supra* note 144, at 45–48.

199. Schragger, *supra* note 166, at 1859 (emphasis added); STIGLITZ, *supra* note 151, at 737.

200. See Allen Erbsen, *Horizontal Federalism*, 93 MINN. L. REV. 493, 522–25 (2008) (discussing negative and positive interstate externalities).

1. *Positive Externalities*

The first type of externality is positive. Knowledge generated in one state may spill over to other states *very quickly*, before the sponsoring state can capture any of the benefits. Without the option for federal funding, states may not fund these “high-spillover” innovations. Section V.B.1 discusses precisely which kinds of innovation are currently funded by the national government. First, states are unlikely to independently fund basic research in fields like physics and anthropology that have little near-term commercial relevance.²⁰¹ Second, states are also unlikely to fund research related to “national public goods,” such as national security or public health, absent the chance to win federal research contracts. In short, one can make the following generalization: states’ incentive to invest in innovation is at its peak when the innovation is likely to engage the private sector and create concrete economic “wins” for the regions: employment, tax revenues, local investment, etc.²⁰²

What should one do when externalities deplete states’ incentives to invest in research that nonetheless has high social value for the national as a whole? The answer, according to Wallace Oates,²⁰³ is that the federal government should provide matching funds. As alluded to, the basic principle of fiscal federalism is that “the lowest level of government encompassing, in a spatial sense, the relevant benefits and costs” should be responsible for financing public goods.²⁰⁴ However, this principle prescribes matching grants from the national government, or whichever government benefits, in cases where “the provision of local services generates benefits for residents of other jurisdictions.”²⁰⁵ This suggests a simple prescription: state innovation finance must be supplemented by

201. See DAN BERGLUND & CHRISTOPHER COBURN, PARTNERSHIPS: A COMPENDIUM OF STATE AND FEDERAL COOPERATIVE TECHNOLOGY PROGRAMS 15 (1995) (“For example, state cooperative technology programs rarely support basic research because its results are so easily diffused before any special benefit can be gained (biotechnology can be an exception to this rule).”).

202. *Id.* (noting that, historically, states’ desire to capture economic benefits from their expenditures on innovation has “drive[n] states toward industrially related technology fields where new products and processes can be readily deployed by companies resident in the state”).

203. Oates’s work on fiscal federalism heavily influenced Cooter and Siegel’s collective action federalism theory. See Cooter & Siegel, *supra* note 16, at 137 n.102 (citing OATES, FISCAL FEDERALISM (1972)).

204. Oates, *supra* note 144, at 1122.

205. *Id.* at 1127.

robust federal funding of high-spillover innovations with little near-term commercial potential.²⁰⁶

The existence of federal funding for research can have a significant impact on the economy of the region in which federal grant recipients locate. For example, following World War II, some regions of California developed their economies thanks to sizeable federal investments in defense research based at U.C. San Diego.²⁰⁷ Therefore, it should not be surprising that today many states fund public research partly to attract federal research grant recipients to the region.²⁰⁸ Section V.B.2 discusses this phenomenon further.

2. *Negative Externalities*

Local governments can also impose *negative* externalities on other regions. There are two types of negative externality problems. The first is distortion: where states' investments in innovation lead to reduced private investments in innovations of national importance. For instance, imagine that Florida subsidizes a boat-hull design that has been in the public domain for years but requires further incentive to achieve successful commercialization. While subsidizing the boat-hull design does not directly harm states other than Florida, it could theoretically lead to more investment in boat-hulls, and less investment in, say, innovations in clean energy—thereby reducing national welfare.²⁰⁹

The second negative externality problem is the zero-sum game: where states become enmeshed in a zero-sum competition for talent, capital, and other scarce inputs to innovation and this competition results not in a net

206. *Id.*; see also Inman & Rubinfeld, *supra* note 144, at 46.

207. See Mary Walshok & Joel West, *Serendipity and Symbiosis: UCSD and the Local Wireless Industry*, in PUBLIC UNIVERSITIES AND REGIONAL GROWTH: INSIGHTS FROM THE UNIVERSITY OF CALIFORNIA 126–52 (Martin Kenney & David Mowery eds., 2014) (discussing the role of defense funding during World War II in the development of the wireless communications industry cluster in San Diego).

208. This raises the possibility of wasteful competition among states for federal research dollars—a problem I discuss in a recent paper. See Hrdy, *Cluster Competition*, *supra* note 185.

209. This was the Supreme Court's concern in *Bonito Boats v. Thundercraft Boats*, where the Court stated that state patents for unpatentable innovations might preclude investments in more innovative subject matter, and consequently decided to preempt state "patent-like rights." See 489 U.S. 141, 150–61 (1989). *But see* Douglas G. Lichtman, *The Economics of Innovation: Protecting Unpatentable Goods*, 81 MINN. L. REV. 693, 713–18 (1997) (challenging the Supreme Court's assumption about the effect of state incentives for unpatentable innovations on innovation).

benefit for all states, but a net loss.²¹⁰ As explained above, one implication of the economic federalism literature is that local governments craft incentives in a highly competitive atmosphere in which mobile residents choose among jurisdictions based partly on the different policies and services available. Local governments may therefore be tempted to design incentives whose main purpose is to lure away firms and talent from other places. For instance, Nevada passed a law in 2014 authorizing \$1 billion in tax breaks for Tesla motor company to build a new facility to test and manufacture a new kind of electric car battery. The explicit purpose of this incentive was to encourage Tesla to locate its factory in Nevada rather than California or Michigan. Many doubt whether Tesla really needed that billion dollars for any reason other than to motivate its selection of Nevada over other states.²¹¹

While this inter-jurisdictional competition can theoretically lead to efficiencies in the form of more effective local policy and propitious clustering of residents with similar preferences,²¹² it can also lead to wasteful spending and purposeless movements from one state to another.²¹³ Consequently, several commentators have asserted that negative impacts on

210. See Peter Enrich, *Saving the States from Themselves: Commerce Clause Constraints on State Tax Incentives for Business*, 110 HARV. L. REV. 377, 377–95 (1996) (arguing that the states have become trapped in a zero sum competition for mobile firms and that this has led them to waste money on tax incentives that have little positive impact on local or national welfare); see also Schragger, *supra* note 166, at 1854 (noting that a questionable assumption of “market-preserving federalism” theory is that local policies geared towards “maximizing local revenue” will “do more than simply move existing economic development from old territories to new ones [and will] instead, induce new economic growth, either in the local jurisdiction or in the nation as a whole”); Kirk Stark & Daniel J. Wilson, *What Do We Know About the Interstate Economic Effects of State Tax Incentives?*, 4 GEO. J.L. & PUB. POL’Y 133 (2006) (examining empirical evidence regarding the effects of state tax incentives on other states).

211. Matthew Wald, *Nevada Woos Tesla Plant in Tax Deal, but Economic Benefits Prompt Debate*, N.Y. TIMES, Sept. 12, 2014, at B1.

212. For example, it could be that the \$1 billion incentive package signaled to Tesla that Nevada was a superior location in which to conduct its electric car manufacturing and that Nevada values Tesla more highly than the other competing states. For this argument, see Clayton P. Gillette, *Business Incentives, Interstate Competition, and the Commerce Clause*, 82 MINN. L. REV. 447, 457–63 (1997) (arguing that state incentives for businesses facilitate efficient competition for scarce resources and helps allocate those resources to the most suitable region that values them most highly).

213. See, e.g., Enrich, *supra* note 210, at 377–81.

the states as a whole justify judicial preemption of state innovation incentives under the Dormant Commerce Clause.²¹⁴

Of course, it's very difficult for economists, let alone a court, to distinguish between illegitimate "raiding" of other states, and a value-adding innovation strategy. As more data is gathered, it could be possible to distinguish between incentives that lead to more productive activity than we would otherwise have and incentives that do nothing more than shift already-productive activity from one state to another.²¹⁵

3. *Immobility and Broken Political Process*

For a variety of reasons—capture by powerful interest groups,²¹⁶ lack of information, corruption, or simple incompetence—local officials may not design high quality, public-spirited innovation policies. Specifically, with respect to innovation finance, they may invest public money in innovations that do not pan out or that require no incentive in the first place. In these situations, the region will not experience economic benefits from the innovation or, if it does, the benefits will not make up for the lost revenues (i.e., this is deadweight loss). Unfortunately, the economic federalism model does not strictly care about the local costs of poorly crafted local innovation policies so long as the effects are confined to the jurisdiction. This is because the economic federalism theories presented above—that local officials have better incentives to craft local innovation policies, and that mobile firms, capital, and talent will sort themselves into efficient clusters—rely on two key assumptions about the local political process.

The first assumption is that residents are effectively mobile and constantly seeking alternative locations with superior policies.²¹⁷ The second assumption is that a functioning democratic process forces local

214. See, e.g., *id.* at 381, 422–67 (arguing that the Dormant Commerce Clause restricts states' authority to use tax incentives as location incentives and urging the Supreme Court to adopt this view).

215. An example of promising new empirical research is Daniel Wilson's work on state R&D tax credits. See, e.g., Daniel J. Wilson, *Beggar Thy Neighbor? The In-State, Out-of-State, and Aggregate Effects of R&D Tax Credits*, 91 REV. ECON. & STAT. 431 (2009); see also Camilla A. Hrdy, *Moretti & Wilson: Do State Innovation Incentives Work?*, WRITTEN DESCRIPTION (Aug. 23, 2013), <http://writtendescription.blogspot.com/2013/08/moretti-wilson-do-state-incentives-for.html>.

216. See DANIEL CARPENTER & DAVID MOSS, PREVENTING REGULATORY CAPTURE: SPECIAL INFLUENCE AND HOW TO LIMIT IT 13–14 (2014) (defining regulatory capture as the result or process by which regulation is consistently directed away from the public interest by the intent of industry itself).

217. Tiebout, *supra* note 173, at 419 (stating that his model is based on the assumption that "[c]onsumer-voters are fully mobile and will move to that community where their preference patterns, which are set, are satisfied").

officials to act in the public interest.²¹⁸ As Weingast puts it, because the mobility of capital and labor limits officials' ability to cater to minority interests, federalism "greatly diminishes the level and pervasiveness of rent-seeking and the formation of distributional coalitions."²¹⁹ The upshot, for our purposes, is that bad innovation policies risk innovators leaving or not entering the state and risk politicians being voted out of power by existing residents.

While I recognize commentators' arguments that these assumptions do not always hold true in practice, it is beyond the scope of this Article to defend or reject the entire economic federalism model.²²⁰ My conception of local governments' role in innovation policy is an ideal to strive for if not always a reality.

4. *A Note on Geographic Inequality*

A very different kind of problem is severe geographic inequality. Economic federalism's main interest is in promoting efficiency in policy design and provisioning of public goods. As just explained, it achieves this goal by assuming that residents' willingness to pay taxes and remain in the jurisdiction will produce those policies in practice. If Kansas needs a new university, Kansas shall have it. If angel investors refuse to come to the jurisdiction unless they can get a tax credit for investments in not-yet-profitable start-ups, the state will offer it, especially if other states are already doing so.

But in reality, not everyone—and not every state—has the same ability to actually pay for what it desires. Different locations across the United States have different constitutions of residents, some with much lower ability to pay for effective innovation policies than others. Kansas may not actually be able to fund a new university and angel tax credits through tax policy or by taking on debt, even if Oklahoma or Massachusetts can. Economic federalism does not strictly care about this problem. So long as inter-jurisdictional competition for residents results in several high-

218. Cooter & Siegel, *supra* note 14, at 138 (asserting that local officials are effectively monitored by local residents); *see also* Weingast, *The Economic Role of Political Institutions*, *supra* note 167, at 5–6 (discussing the limits the threat of flight places on political actors).

219. Weingast, *The Economic Role of Political Institutions*, *supra* note 167, at 6.

220. *See* Schragger, *supra* note 161, at 1853–54 (critiquing Weingast's assumption that competition limits predation); *see also* David Schleicher, *Federalism and State Democracy*, 95 TEX. L. REV. (forthcoming 2017) (noting that each of the common justifications for federalism requires "state democracy to actually function"), draft available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2739791###.

performing innovation clusters across the country in different technology areas, it does not matter whether there are losers in this race.

As discussed in Section II.C, the United States is already to some degree divided into “dominant clusters” that “continually pull in firms, entrepreneurs and workers,” and “lower tier regions” that find it “difficult . . . to break into the dominant groups.”²²¹ Should the federal government start to intervene? At what point do we decide as a national taxpaying body that—regardless of efficiency—this division between “dominant” and “lower tier” regions results in a level of geographic inequality that is socially undesirable?

I address these difficult questions in a separate article, where I show that beginning around 2010 the federal government has begun to intervene in the local competition to grow clusters by providing subsidies and other opportunities to help selected regions craft effective innovation policies.²²² For example, in 2010, the America COMPETES Reauthorization Act authorized a new regional innovation program to assist state and local governments in developing clusters in selected technology areas. The Act explicitly gives preference to underdeveloped regions negatively impacted by trade.²²³ I argue that federal involvement in cluster competition can be justified based both on distributional concerns and based on efficiency. Not unlike the patent system, federal subsidies for regions to innovate in selected technology areas can theoretically prevent duplicative investments in research and technology development and limit rent-dissipating competition to be the next Silicon Valley.

V. FEDERAL AND STATE INNOVATION POLICY IN PRACTICE

Part IV addressed how theories of economic federalism suggest U.S. innovation policy *ought* to be structured. Collective action federalism and its guiding internalization principle mandate a different conclusion depending on which type of innovation policy we select: intellectual property rights, on the one hand, or innovation finance, on the other. If patents are selected, jurisdiction is more likely to be national or even global due to the problem of spillovers. But if innovation finance is selected,

221. See Karen G. Mills, Andrew Reamer & Elisabeth B. Reynolds, *Clusters and Competitiveness: A New Federal Role for Stimulating Regional Economies*, BROOKINGS INST. (Apr. 2008), at 12, <https://www.brookings.edu/wp-content/uploads/2016/07/Clusters-Brief.pdf>.

222. See Hrdy, *Cluster Competition*, *supra* note 185.

223. See America COMPETES Reauthorization Act of 2010, Pub. L. 111-358, § 603, 124 Stat. 3982, 4030–37 (2011) (codified as amended at 15 U.S.C. § 3722 (2012)).

jurisdiction should often be local since the economic benefits of innovation are highly localized to the region in which researchers, firms, and skilled workers locate.

This final Part is largely descriptive. I show that the internalization principle—which assigns authority to the smallest jurisdiction that internalizes the benefits of its policies and reserves national action for collective action failure—is a remarkably good descriptor for how federal and state governments actually respond to the problem of promoting innovation in their respective jurisdictions. Patent law is purely federal law and has become increasingly so over the centuries. Innovation finance is offered concurrently by national and local governments. With respect to commercial innovations for which there is already a private market and the potential for localized benefits is high, the innovation finance options are far more extensive at the state and local level.

A. PATENT LAW

Collective action federalism theory works very well for patent law. As I show below, patent law's jurisdictional trajectory from state to federal rights seems to have been explicitly motivated by the problem of spillovers: cross-border copying of inventions protected in other states. Protecting patents, along with copyrights, may well represent a collective action failure that necessitates federal intervention under the IP Clause, just as Cooter and Siegel contend.²²⁴ On the other hand, states' retrenchment from patent law stands in stark contrast with states' historic and continuing role in protecting businesses' trade secrets from misappropriation.

1. *From State to Federal Rights*

Prior to ratification of the Constitution in 1788, U.S. states had their own patent laws.²²⁵ However, the advent of interstate commerce challenged the efficacy of state-level patent regimes. Inventors ceased to be able to effectually protect their rights within a free-trading national marketplace.²²⁶ When the time came to draft a Constitution laying out the powers of the new federal government, no one appeared to object to James Madison's famous statement in *The Federalist* No. 43 that "[t]he States cannot separately make effectual provision for either . . . [patent or copyright] cases."²²⁷ Thereafter, the Framers added the IP Clause to the Constitution with little discussion of

224. Cooter & Siegel, *supra* note 16, at 148–49.

225. See Hrdy, *State Patent Laws in the Age of Laissez Faire*, *supra* note 25, at 58–67.

226. *Id.* at 67–70.

227. THE FEDERALIST NO. 43, at 279 (James Madison) (Clinton Rossiter ed., 1961).

the matter. In 1790 and 1793, two federal patent acts passed in rapid succession.²²⁸ A few states, such as New Hampshire and New Jersey, continued granting patents after ratification, but the Patent Act of 1793 clarified that these patents would be automatically relinquished upon obtaining a federal patent.²²⁹

In 1836, Congress created a national Patent Office to review applications and issue granted patents.²³⁰ The creation of a national Patent Office appeared to complete the federalization of the substance of patent law. Under the post-1836 system, writes Herbert Hovenkamp, “[t]he federal patent . . . evolved into a ‘property right’ that applicants could obtain through an administrative procedure intended to be politically neutral, and that patentees could practice or not at their will.”²³¹ Meanwhile, Hovenkamp recounts, states’ role in the creation of property rights in inventions, as compared to corporate charters or other exclusive franchises, appeared minimal. Even as the granting of exclusive rights in corporate charters “remained largely a function of the states . . . the power to grant exclusive rights for inventions came to be seen as a federal prerogative.”²³²

This trend towards patent federalization makes sense from the perspective of the efficiency values underlying collective action federalism. The fundamental problem was that patent disclosures produced a severe inter-jurisdictional externality: benefits for other states at the expense of the state that offered the patents and copyrights.²³³ Prior to the existence of national patents, creators were forced to apply for rights in all states in which they sought to market their invention or enforce their rights. This was

228. See Patent Act of 1790, ch. 7, 1 Stat. 109 (1790) (repealed 1793); Patent Act of 1793, ch. 11, 1 Stat. 318 (1793) (repealed 1836).

229. Patent Act of 1793, ch. 11, § 7, 1 Stat. 318, 322 (1793); see Hrdy, *State Patents in the Age of Laissez Faire*, *supra* note 25, at 72–74, 77.

230. See Patent Act of 1836, ch. 357, § 18, 5 Stat. 117, 124 (1836) (repealed 1861).

231. Herbert Hovenkamp, *The Emergence of Classical Patent Law in American Legal Thought*, 58 ARIZ. L. REV. 263, 270 (2016).

232. *Id.* at 278. Corporate charters were different from federal patents because they created exclusive rights to operate an enterprise, such as a bridge or a toll road, in a physical space. On state monopolies in the nineteenth century, see Herbert Hovenkamp, *Technology, Politics, and Regulated Monopoly: An American Historical Perspective*, 62 TEX. L. REV. 1263, 1268 (1984).

233. See Hrdy, *State Patents as a Solution*, *supra* note 20, at 83 (noting that “state patents would . . . generate valuable ‘innovation spillovers’ for innovators and competitors across the country”); see also EDWARD WALTERSCHEID, *THE NATURE OF THE INTELLECTUAL PROPERTY CLAUSE: A STUDY IN HISTORICAL PERSPECTIVE* 76 (2000) (“The most singular defect was that states could only legislate with respect to their own territory.”).

costly from the perspective of rights owners, and it was risky from the perspective of states.²³⁴

The problem could potentially have been resolved through inter-state bargaining. But this solution faced significant transaction costs, requiring an agreement among up to fifty states, each of which had an incentive to cheat by continuing to allow copying of the productions of their neighbors. As explained in Part IV, economic federalism theory favors nationalizing when the transaction costs implicated by inter-state bargaining are very high.²³⁵ Thus, patents and copyrights should, at a substantive level, generally be national rights to avoid the inefficiencies of bargaining between states.²³⁶

That said, as I have previously shown, there are nonetheless instances in which states could productively grant patents for geographically localized innovations, such as advancements in agricultural technology, water conservation, or energy production. Unlike in copyright, patentable subject matter often has significant geographic dimensions and utility can vary from place to place.²³⁷ Markets for inventions may be limited to the region, as in the case of some agricultural innovations. Because commercialization in the patent context is difficult, the availability of a U.S. patent alone does not always ensure practice and implementation; a state patent could improve a developer's incentive and ability to undertake local

234. See Hrdy, *State Patent Laws in the Age of Laissez Faire*, *supra* note 25, at 67–70; WALTERSCHEID, *supra* note 233, at 76 (“Getting multiple state patents or copyrights was time consuming, expensive, and frequently frustrating.”).

235. Cooter & Siegel, *supra* note 16, at 139–40 (discussing the “Federal Coase Theorem”).

236. Efficiency also likely weighs against giving state courts concurrent authority to make determinations of patent validity and infringement, though an analogy could be drawn to the dual system we currently have in which the Patent Trial and Appeals Board (PTAB) makes determinations of validity alongside federal courts. See Ben Picozzi, Comment, *Finality in Parallel Patent Proceedings*, 125 YALE L.J. 2519 (2016). On the gradual demise of state courts' concurrent authority to entertain patent lawsuits, see Paul Gugliuzza, *Patent Law Federalism*, 2014 WIS. L. REV. 11, 17–19 (2014). *But see id.* at 27–75 (challenging the prevailing assumption that federal courts should have exclusive jurisdiction in cases arising under patent law); Edward Cooper, *State Law of Patent Exploitation*, 56 MINN. L. REV. 313, 318–24, 344–73 (1972) (discussing various types of cases where state courts have historically been called upon to determine patent-related issues arising under tort and contract, and even to determine the scope and validity of patents).

237. See *Goldstein v. California*, 412 U.S. 546, 556–57 (1973) (“The patents granted by the States in the 18th century show, to the contrary, a willingness on the part of the States to promote those portions of science and the arts which were of local importance.”).

commercialization of high cost, high-risk ventures.²³⁸ Thus, I have previously argued that in certain circumstances states should be able to grant their own domestic patent rights in exchange for the promise to commercialize innovations of high local utility to the jurisdiction.²³⁹ The result should be more commercialization of socially valuable localized technologies as well as potentially useful experiments in patent law and policy.²⁴⁰

2. *A Contrast with Trade Secrets*

The internalization principle comes to a very different result for forms of intellectual property that do not implicate broad disclosure of new and easily transferable information. Patents correct for the externalities involved in innovation by creating exclusive rights, but they do so only for new and nonobvious inventions *that are fully disclosed in a patent document*.²⁴¹ Trade secrets, in contrast, protect information that is by definition nonpublic—that is, information that has never been disclosed to the public at large, and that has been kept secret using reasonable efforts—and they only protect against “improper” acquisition, use, or disclosure of that information.²⁴² Thus, unlike patents, trade secrets do not inevitably implicate severe inter-state spillover of easily transferable information. To

238. Hrdy, *State Patents as a Solution*, *supra* note 20, at 116–19 (discussing the potential role of a state patent in promoting commercialization).

239. *Id.* at 102–103.

240. *Id.* at 103 (“Especially given the continued move towards global uniformity, patent law could benefit from the policy experiments that divergent state patent regimes would produce, turning the states into decentralized ‘laboratories’ for improving the functioning of patent law.”).

241. Patent law’s longstanding disclosure requirement has two functions: to “permit society at large to apply the information by freely making or using the patented invention after the expiration of the patent,” and to “stimulate others to design around the invention or conceive of new inventions—either by improving upon the invention or by being inspired by it—even during the patent term.” Jeanne Fromer, *Patent Disclosure*, 94 IOWA L. REV. 439, 458–59 (2009).

242. See generally UNIF. TRADE SECRETS ACT § 4 (amended 1985), 14 U.L.A. 536–59 (2005); see also *Kewanee v. Bicron*, 416 U.S. 470, 484 (1974) (“By definition a trade secret has not been placed in the public domain.”); *E. I. duPont deNemours & Co. v. Christopher*, 431 F.2d 1012, 1016 (5th Cir. 1970) (applying Texas law) (“DuPont has a valid cause of action to prohibit [the defendant] from improperly discovering its trade secret and to prohibit the undisclosed third party from using the improperly obtained information.”). *But see* Michael Risch, *Hidden in Plain Sight*, 31 BERKELEY TECH. L.J. (forthcoming 2017) (asserting that in actuality “[a] long line of cases—in virtually every circuit—provides for the protection of trade secrets in products sold to the public”), draft available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2761100.

the contrary, while patents *promote* disclosure of information, trade secrets *prevent* disclosure of information.

In line with this prediction, trade secret law is the only IP regime that remains primarily state law.²⁴³ Trade secret laws originated in disparate state common law rules.²⁴⁴ They were eventually made more uniform beginning with the American Law Institute's (ALI) *Restatement (First) of Torts* (1939) and culminating with the Uniform Trade Secrets Act (UTSA) in 1979, which has been adopted in forty-seven states.²⁴⁵ The Supreme Court has held that states wishing to use trade secret laws to provide incentives for non-patentable subject matter can do so, as long as they limit protection to innovations that have not yet been publicly disclosed and do not prevent reverse engineering.²⁴⁶

This allocation of power has now shifted. Congress has passed a federal *civil* cause of action for trade secret misappropriation under federal law.²⁴⁷ The major function of the Defend Trade Secrets Act (DTSA) is to broaden the scope of trade secret protection in actions involving extraterritorial conduct and to facilitate enforcement in situations involving multiple states.²⁴⁸ The driving motivation is clear: to provide a more efficient

243. See Robert Denicola, *The Restatements, the Uniform Act and the Status of American Trade Secret Law*, in *THE LAW AND THEORY OF TRADE SECRECY* 18 (Rochelle C. Dreyfuss & Katherine J. Strandburg eds., 2011) (noting that unlike patent, copyright, and trademark law, “[t]rade secret law, however, is state law”).

244. *Id.* at 18–22; see also ELIZABETH ROWE & SHARON SANDEEN, *CASES AND MATERIALS ON TRADE SECRET LAW* 14–15 (1st ed. 2012).

245. See Denicola, *supra* note 243, at 20–21, 33–44 (discussing the UTSA’s adoption and its effect on preexisting common law protections).

246. As the Court observed in *Kewanee v. Bicron* (1974), where it rejected a challenge to the constitutionality of state trade secrets laws under the Supremacy Clause, there is no inherent reason states cannot perform these functions for the benefit of industry. “Trade secret law,” the Court stated, “encourages the development and exploitation of those items of lesser or different invention than might be accorded protection under the patent laws Until Congress takes affirmative action to the contrary, States should be free to grant protection to trade secrets.” 416 U.S. at 493.

247. Defend Trade Secrets Act of 2016, Pub. L. No. 114-153, 130 Stat. 376 (2016) (to be codified at 18 U.S.C. §§ 1832(b), 1833, 1835, 1836(b)-(d), 1839(3)-(7)); see also Christopher Seaman, *The Case Against Federalizing Trade Secrecy*, 101 VA. L. REV. 317, 320 (2015) (arguing federalization is unnecessary in light of state protection and criminal liability under the EEA); David Levine & Sharon Sandeen, *Here Come the Trade Secret Trolls*, 71 WASH. & LEE L. REV. ONLINE 230 (2015) (arguing federalization risks spawning new forms of “trolling behavior”). *But see* James Pooley, *The Myth of the Trade Secret Troll: Why We Need a Federal Civil Claim for Trade Secret Misappropriation*, 23 GEO. MASON L. REV. 1045 (2016) (arguing that national trade secret protection is both necessary and efficient).

248. See Seaman, *supra* note 247, at 340–48.

mechanism to keep knowledge generated in U.S. borders in the country, and to enable individual states to more effectively protect the trade secrets of local companies.²⁴⁹

Federalization was not without its critics. For example, Christopher Seaman argues that federalization was unnecessary, given that state trade secret laws already exist and are nearly uniform; moreover, parties can already access a federal forum in cases involving citizens from different states, including foreign states.²⁵⁰ Seaman also worries that federalizing trade secret law will make trade secret protections far stronger as compared to patent law, leading to an increase in secrecy and a reduction in patenting and public disclosure.²⁵¹ Lastly, Seaman suggests that states should be allowed some flexibility to experiment in trade secret law.²⁵²

Collective action federalism provides a more fundamental reason for objecting to complete federalization. Rather than asking whether a national trade secret law would make trade secret protection weaker or stronger, the theory looks instead at what would be the most efficient allocation of government. In this case, the knowledge spillovers potentially created by the production trade secret information within a firm will often be localized. Even in the digital age, most trade secret theft is performed by insiders, “typically involving alleged breaches of confidence in the context of business-to-business and employer–employee relationships.”²⁵³ Many of these thefts will result in involuntary knowledge transfers—but only to other employees or companies in the region. Even in cases where employees leave a state for another state within the U.S., there is no inherent reason that state or federal courts cannot obtain jurisdiction over the departing employee, just as they do in ordinary tort cases.²⁵⁴

249. See Pooley, *supra* note 247, at 3 (“[F]ederalizing trade secret law would fill a critical gap in effective enforcement of private rights against cross-border misappropriation that has become too stealthy and quick to be dealt with predictably in state courts.”).

250. Seaman, *supra* note 247, at 369–70 (noting that diversity jurisdiction covers cases brought by U.S. corporations against citizens of foreign states, “which is particularly useful in cases involving trade secret misappropriation by foreign entities”).

251. *Id.* at 379–85; see also Camilla A. Hrdy, *Seaman: The Case Against Federalizing Trade Secrecy*, WRITTEN DESCRIPTION (Apr. 30, 2014), <http://writtendescription.blogspot.com/2014/04/seaman-case-against-federalizing-trade.html>.

252. Seaman, *supra* note 247, at 322 (“[T]here are benefits to a decentralized approach that permits states to engage in a limited degree of experimentation regarding the scope of trade secret protection.”).

253. Levine & Sandeen, *supra* note 247, at 239–40 (“The existing data establishes that the bulk of all trade secret cases are of the domestic variety . . .”).

254. See, e.g., *Clorox v. S.C. Johnson & Son, Inc.*, 627 F. Supp. 2d 964, 972 (E.D. Wis. 2009) (denying motion for temporary restraining order and injunction to prevent an

There is one very important exception where states cannot internalize the benefits of their trade secrets laws: when someone misappropriates information leaves the state's jurisdiction—taking the trade secret with them. For example, if a firm in New Jersey hires employees from Taiwan, and these employees leave the country in possession of the firm's trade secrets, New Jersey may have difficulty obtaining jurisdiction over the defendants or obtaining a remedy under Taiwanese law. Unless the state can obtain extraterritorial jurisdiction, this represents a positive spillover to a non-U.S. country that state law is not competent to address.²⁵⁵ Depending on the transaction costs of bargaining with other countries to expand extraterritorial protection for trade secrets, the collective action principle suggests that nationalization is appropriate *in the specific case of foreign espionage*.

This too tracks reality. Even before the DTSA was passed, the United States had a criminal remedy in place for misappropriations that implicate cross-border spillovers. As Elizabeth Rowe and Daniel Mahfood have discussed, “[i]nternational espionage of American trade secrets” has become a “growing problem with wide-ranging significance implicating national security, economic, and political interests.”²⁵⁶ In response, Congress passed the Economic Espionage Act (EEA) in 1996,²⁵⁷ primarily to confront the threat of foreigners leaving the United States with the secret information of American companies.²⁵⁸ That said, the EEA allows only criminal causes of action, and the decision to prosecute is subject to the discretion of U.S. Attorneys. Thus, the EEA remained “a limited option for private companies.”²⁵⁹

employee based at a California company from revealing confidential information in his new employment at a Wisconsin company).

255. See Elizabeth Rowe & Daniel Mahfood, *Trade Secrets, Trade, and Extraterritoriality*, 66 ALA. L. REV. 63, 64 (2014) (“A primary obstacle to [protecting trade secrets abroad] is the principle of territoriality—the notion that U.S. law applies only to acts that take place on U.S. soil. As a consequence of this principle, American companies doing business abroad, or whose trade secrets are misappropriated abroad, have limited recourse against a potential infringer through either criminal or civil actions.”).

256. *Id.* at 67.

257. Pub. L. No. 104-294, 110 Stat. 3488 (1996) (codified as amended at 18 U.S.C. §§ 1831-1839 (2012)).

258. The Act criminalizes espionage on behalf of a foreign entity and theft of trade secrets for pecuniary gain and also provides a criminal remedy for thefts that involve products sold in interstate commerce. Unlike state trade secret laws, the EEA applies to extraterritorial conduct by U.S. citizens and non-citizens and clearly targets theft involving either flight to a foreign country or action by foreign entities. 18 U.S.C. §§ 1831, 1832 (2012); see also Rowe & Mahfood, *supra* note 255, at 64, 102.

259. Rowe & Mahfood, *supra* note 255, at 64.

This limitation was not a flaw. As a matter of efficient allocation of government, it was appropriate to have a limited, targeted federal remedy to confront cross-border thefts that challenged state courts' jurisdiction and federal district courts' jurisdiction in diversity cases.²⁶⁰ Given that there was no pervasive collective action failure requiring federal law in *every case* or even in the average case, economic federalism suggests that states should remain responsible for trade secret laws to obtain the benefits of local jurisdiction discussed above.²⁶¹

Along with the desirability of some legal experimentation, discussed by Seaman, the theories discussed in Section III.C suggest two other reasons to retain state jurisdiction and allow for variation among the states when variation occurs.²⁶² First, according to the “market-preserving” federalism theory explained in Section III.C.1, states may simply have better incentives and better information to design trade secret laws that match the needs of specific industries and people within their jurisdictions. For example, one state may decide that intrastate norms and practices support strong trade secret protection when an employee has threatened to leave a firm but has not yet used or disclosed any secrets; other states might favor weaker trade secret protection in such circumstances.²⁶³ Second, according to the theory of Tiebout clustering explained in Section III.C.2, firms and industries can choose which rule works better for them, leading to more efficient matching of firms to different jurisdictions and more efficient production within those jurisdictions and overall.

260. Absent express authority from Congress, federal courts, like state courts, are limited by the long-arm statutes of the states in which they are located. *See* FED. R. CIV. P. 4(k)(1)(A) (stating that jurisdiction over defendants is proper where the defendant could be subjected to the jurisdiction of state courts in the state where the federal district court is located).

261. *See, e.g.*, Yoo, *supra* note 161, at 1402–03; *see also* Section III.C.

262. Note that even though most states have adopted the UTSA, there are still significant differences between states both statutory and common law. *See* Pooley, *supra* note 247, at 5–6 (discussing the lack of uniformity even within UTSA states and giving examples of divergent rules among the states).

263. California courts reject the “inevitable disclosure doctrine” on the grounds that it interferes with California’s policy of free mobility and would create an “after-the-fact covenant not to compete” restricting employee mobility. *See, e.g.*, *Les Concierges, Inc. v. Robeson*, No. C-09-1510-MMC, 2009 WL 1138561 (N.D. Cal. Apr. 26, 2009). Pennsylvania courts, in contrast, allow protection in such circumstances. *Insulation Group LLC v. Sproule*, 613 F. Supp. 2d 844, 855 (S.D. Tex. 2009) (applying Pennsylvania law to grant an injunction without any evidence that the former employee had actually used or disclosed the former employer’s trade secrets because it is “impossible for [the former employee] not to disclose [the] trade secrets”).

The possibility of such benefits from localization of authority suggests that, since states *can* internalize the effects of prohibiting trade secret misappropriation within their own borders in many if not all cases, general civil trade secret law should remain state law. On the other hand, the new system takes a weak compromise, creating a federal civil trade secret law while declining to preempt state trade secret law.²⁶⁴ There seems little reason to give litigants the option to bring both federal and state law trade secret claims when state law would be sufficient on its own. This seems a wasteful duplication of remedies.

B. INNOVATION FINANCE

Thus, to sum up, patent law and now trade secret law have become federal law. Cross-border spillovers—that is, the risk of out-of-state copying or involuntary transfer of knowledge from one location to another—have been the main culprit. With respect to innovation finance, however, the situation is very different. Unlike IP rights, direct public financing of innovation does not operate through the mechanism of exclusion, and thus does not suffer from the same problem.²⁶⁵

This insight prevents the simple conclusion that innovation finance, like patents, should generally be national. Instead, the smallest jurisdiction that internalizes the benefits of innovation finance should be responsible. As explained in Section II.C, the near term, and even the long-term, benefits of innovation tend to be concentrated in the region in which innovators locate their research and operations and in which they employ workers at high-skill wages. In such cases, innovation finance should be the responsibility of the subnational region that actually experiences these benefits. There are exceptions to this rule. As noted in Section IV.D.1, the main exception is “high-spillover” research that does not create sufficient local economic benefits to justify local funding without the chance for national supplements.

In the sections below, I demonstrate that this prediction is borne out in practice by performing an in-depth investigation of the innovation incentives actually provided in the United States today at the federal and state levels.

264. The DTSA does not preempt state law. Pub. L. No. 114-153, § 2, 130 Stat. 376, 381 (2016) (to amend 18 U.S.C. §§ 1833, 1836 (2012)) (“Nothing in the amendments made by this section shall be construed to modify the rule of construction under section 1838 of title 18, United States Code, or to preempt any other provision of law.”).

265. *See supra* Section IV.B.

1. U.S. Federal Innovation Finance

Innovation in the United States is driven primarily by the private sector. The total R&D performed for 2010-11 was over \$400 billion. The private sector conducted around \$284 billion of this amount, and funded around \$248 billion of the total.²⁶⁶ However, the U.S. federal government funds a significant amount of R&D every year. After World War II, the federal government significantly increased its share of funding for science and technology-based research.²⁶⁷ Today the national government spends around \$130 billion a year on research and development, around 30% of the national total.²⁶⁸

However, the cases where the federal government directly finances research are the exception rather than the rule. Federally funded research is limited to those cases in which innovation produces such significant national benefits that states alone are not willing to fund it: basic science research with no market or known commercial relevance; R&D in specific national mission areas; and “dual use” funding for private enterprises innovating in those areas.²⁶⁹ This Article argues that each of these circumstances entails significant national spillovers that justify at least some federal financing *despite* the fact that many of the economic benefits will be localized to the regions in which the research is performed.

266. See Mark Boroush, *U.S. R&D Spending Resumes Growth in 2010 and 2011 but Still Lags Behind the Pace of Expansion of the National Economy*, NSF: INFO BRIEF, NAT'L CTR. SCI. & ENG'R STATISTICS (NSF 13-313, Jan. 2013), <http://www.nsf.gov/statistics/infbrief/nsf13313/nsf13313.pdf>; see also Mark Boroush, *National Patterns of R&D Resources: 2010-2011 Data Update*, NSF: NAT'L CTR. SCI. & ENG'R STATISTICS (NSF 13-318, Apr. 2013), <http://www.nsf.gov/statistics/nsf13318/pdf/nsf13318.pdf>. For discussions of the private sector as the main driver of innovation, see, for example, LERNER, *supra* note 27, at 150; SCOTCHMER, *supra* note 6, at 140-43; Keller, *supra* note 28, at 110-11.

267. LERNER, *supra* note 27, at 33, 20, 150-52.

268. NSF: NAT'L CTR. SCI. & ENG'R STATISTICS, FEDERAL FUNDS FOR RESEARCH AND DEVELOPMENT FYS 2012-14, tbl. 3, “Federal obligations for research, development, and R&D plant, by agency,” <https://www.nsf.gov/statistics/nsf14316/pdf/tab3.pdf>. For an overview of federal funding for R&D, see David Mowery & Richard Rosenberg, *The U.S. National Innovation System*, in NATIONAL INNOVATION SYSTEMS: A COMPARATIVE ANALYSIS 29-75 (Richard Nelson ed., 1993); Fred Block, *Innovation and the Invisible Hand of Government*, in STATE OF INNOVATION: THE U.S. GOVERNMENT'S ROLE IN TECHNOLOGY DEVELOPMENT 4-30 (Fred Block & Matthew R. Keller eds., 2011).

269. BRANSCOMB & AUERSWALD, *supra* note 28, at 144. On the federal government's mission-oriented technology policy, see Lewis M. Branscomb & George Parker, *Funding Civilian and Dual-Use Industrial Technology*, in EMPOWERING TECHNOLOGY: IMPLEMENTING A U.S. STRATEGY 64, 69 (Lewis M. Branscomb ed., 1993); Maryellen Kelley, *From Core Mission to Commercial Orientation: Perils and Possibilities for Federal Industrial Technology Policy*, 11 ECON. DEV. Q. 313, 315-26 (1997).

a) Basic Science Research

Basic science research is research that has no known practical, commercial, or government application.²⁷⁰ It is difficult for states to internalize the full value of basic science research because the resulting knowledge can be used by others outside the state either instantly or in the near future *without first creating localized economic benefits like employment or new firms*. Thus, the collective action failure principle suggests that funding must come from the national government rather than subnational governments.

This prediction holds true in practice. While states do fund some basic research—a phenomenon discussed in the next section—basic research with no defined commercial application relies heavily on federal subsidies. In 1950, Congress created the National Science Foundation (NSF), an independent federal agency whose mission is “to promote the progress of science; to advance the national health, prosperity, and welfare; to secure the national defense”²⁷¹ The NSF funds much of the basic research (around 24%) performed at universities and is the only federal agency dedicated to supporting fundamental research and education in all scientific and engineering disciplines, except the medical sciences.²⁷² Supported research areas include the Biological Sciences, Mathematical & Physical Sciences, and Social, Behavioral & Economic Sciences.²⁷³ The NSF has an annual budget of around \$7.5 billion a year.²⁷⁴

b) Mission R&D

Besides basic research, the federal government funds R&D in selected subject matter areas relevant to national missions. The federal government administers most federal research money through national mission agencies with mandates to focus on specific research areas of high national relevance. Today the major mission agencies that fund research are the Department of Defense (DOD), the Department of Health and Human Services (HHS) (home to the National Institutes of Health (NIH)), the National Aeronautics

270. Sohvi Leih & David J. Teece, “Basic Research,” THE PALGRAVE ENCYCLOPEDIA OF STRATEGIC MANAGEMENT (Mie Augier & David J. Teece eds., 2016) (defining the term).

271. NSF, ABOUT THE NATIONAL SCIENCE FOUNDATION, <https://www.nsf.gov/about/> (last visited Feb. 9, 2016).

272. *Id.*

273. NSF, FIND FUNDING, <https://www.nsf.gov/funding/index.jsp#areas> (last visited Feb. 9, 2016).

274. NSF, ABOUT THE NATIONAL SCIENCE FOUNDATION, <https://www.nsf.gov/about/> (last visited Feb. 9, 2016).

and Space Administration (NASA), the Department of Energy (DOE), the Department of Transportation (DOT), and the Department of the Interior (DOI). In addition, the National Institute for Standards and Technology (NIST), a relatively small agency in the Department of Commerce (DOC), receives around \$850 million a year for programs oriented towards research and commercialization.²⁷⁵ Federal R&D grants go to four main actors: private businesses; federal agencies, which perform intramural R&D in their own research facilities;²⁷⁶ universities and colleges; and nonprofits.²⁷⁷

The table below lists the top four agencies through which the national government funds basic and applied research, and the federal funds provided to each agency in 2012. Of the approximately \$130 billion that the federal government spends per year on research, *about half* goes towards defense research and domestic security. Health research comes in second, receiving around \$30 billion a year.²⁷⁸

Table 1: Federal Funding for R&D in Top Four Mission Areas in 2012 (in billions)²⁷⁹

DOD	\$65.3
HHS	\$30.9
NASA	\$10.6
DOE	\$10.3

275. See NIST APPROPRIATIONS SUMMARY FY 2013 – FY 2015, <https://www.nist.gov/director/congressional-and-legislative-affairs/nist-appropriations-summary-fy-2013-fy-2015>; see also JOHN F. SARGENT JR., CONG. RESEARCH SERV., R43580, FEDERAL RESEARCH AND DEVELOPMENT FUNDING: FY 2015 5–6 tbl. 1, 48–51, tbl. 15 (2015) (providing recent data on federal funding for R&D and NIST specifically).

276. Rochelle Dreyfuss, *Tailoring Incentives: A Comment on Hemel and Ouellette’s Beyond the Patents–Prizes Debate*, 92 TEX. L. REV. 131, 132 (2014) (“R&D spending by federal laboratories, such as the National Institutes of Health, is substantial. In 2009, for example, intramural spending amounted to \$30.9 billion and constituted 8% of all U.S. R&D expenditures.”).

277. See Boroush, *U.S. R&D Spending Resumes Growth in 2010 and 2011*, *supra* note 266, at 2, 4.

278. Cf. Adam Marcus & Ivan Oransky, *Getting the Bogus Studies Out of Science*, N.Y. TIMES, Aug. 20, 2015, at A11 (noting proposal to increase NIH funding for cancer research by \$9.3 billion).

279. NSF, NAT’L CTR. SCI. & ENG’R STATISTICS, FEDERAL FUNDS FOR RESEARCH AND DEVELOPMENT FYS 2012–14, tbl. 3, “Federal obligations for research, development, and R&D plant, by agency,” <https://www.nsf.gov/statistics/nsf14316/pdf/tab3.pdf>.

c) “Dual Use” Funding

The federal government does fund private, commercial innovations. However, when the government funds such research, it requires the research to have “dual use”: demonstrated commercial relevance *and* relevance to a core federal mission area like defense, health, or efficient energy use.²⁸⁰ There are currently a variety of dual-use programs through which private companies can obtain funding so long as they are conducting research in a national mission area that suits an agency’s needs—especially in the military context.²⁸¹

A prominent example of a dual-use program is the Defense Advanced Research Projects Agency (DARPA), whose stated mission is “to make pivotal investments in breakthrough technologies for national security.”²⁸² DARPA spends over \$2 billion a year on projects such as advanced humanoid robots.²⁸³ Another high-profile example is In-Q-Tel, the Central Intelligence Agency’s venture capital arm that finances innovation related to CIA missions, and sometimes obtains equity stakes in the companies it funds.²⁸⁴ The most general-purpose dual use program is the Small Business Innovation Research (SBIR) program, which is oriented toward helping small businesses while pursuing national mission goals.²⁸⁵ SBIR’s mandate requires all federal agencies spending over \$100 million annually on extramural research contracts to reserve a portion of their budgets for contracts with for-profit small businesses with under 500 employees.²⁸⁶

280. See Branscomb & Parker, *supra* note 269, at 69; Kelley, *supra* note 269, at 315–26.

281. See Kelley, *supra* note 269, at 317–18 (discussing adoption of dual use programs during the Clinton era and suggesting a broader definition of “dual use” to incorporate non-military investments in mission areas with potential commercial relevance).

282. See DARPA, MISSION, <http://www.darpa.mil/about-us/mission> (last visited Feb. 7, 2017).

283. See DARPA, BUDGET, <http://www.darpa.mil/about-us/budget> (last visited Feb. 7, 2017); DARPA, OUR RESEARCH ARCHIVE, <http://www.darpa.mil/archive/our-research> (last visited Feb. 7, 2017).

284. See Keller, *supra* note 28, at 109–32 (discussing federal venture capital programs for firms developing technologies related to defense and national security); see also John Markoff, *Pentagon Shops in Silicon Valley for Game Changers*, N.Y. TIMES, Feb. 27, 2015, at A3 (discussing the Pentagon’s announcement of a new venture financing program and noting that some companies see committing to do research by the military as problematic).

285. See Small Business Innovation Development Act of 1982, Pub. L. No. 97-219, 96 Stat. 217 (1982) (codified at 15 U.S.C. § 638 (2012)) (re-authorized through 2017 in the National Defense Authorization Act for Fiscal Year 2012, Pub. L. No. 112-81, § 5101, 125 Stat. 1298, 1824 (2011)); 15 U.S.C. § 638(e)(4) (defining SBIR program).

286. See 15 U.S.C. § 638 (f)(1) (listing required set-aside percentages, which increase each fiscal year through 2017).

Funded research must fall into agency mission areas. In practice, most awards go to companies working on technology related to defense and public health.²⁸⁷

In a recent influential book, Mariana Mazzucato argues that, contrary to common perception, the U.S. government directly finances innovation, and should do so more.²⁸⁸ However, despite Mazzucato's characterization of the U.S. government as an "entrepreneurial state," the national government rarely departs from this "dual use mandate."²⁸⁹ When it does, the federal government is accused of engaging in "industrial policy" and "picking winners."²⁹⁰ An example is the Department of Commerce's short-lived venture capital program. For several years, NIST was authorized to operate the Advanced Technology Program (ATP), which provided open-ended equity financing for companies involved in commercializing high-risk research: "In essence, the mission of the . . . [ATP was] to support private sector R&D projects that offer[ed] potential for contributing to technical advance and for realizing economic value."²⁹¹ But the ATP was abolished in 2011 in the midst of wrangling over the national budget.²⁹² The ATP's demise is not an isolated incident. More recently, for instance, pundits

287. In 2012, most SBIR awards were granted through the DOD and the NIH. Small Bus. Innovation Research, *SBIR Annual Report*, SBIR, <https://www.sbir.gov/awards/annual-reports> (last visited Feb. 9, 2016).

288. See MARIANA MAZZUCATO, *THE ENTREPRENEURIAL STATE: DEBUNKING PUBLIC VS. PRIVATE SECTOR MYTHS* (2014); see also Eduardo Porter, *Public R&D, Private Profits and Us*, N.Y. TIMES, May 27, 2015, at B1, B7 (discussing Mazzucato's argument and raising concerns about government taking an equity stake in companies it funds).

289. Mazzucato's characterization of the U.S. government as an "entrepreneurial state" is somewhat misleading. In fact, the U.S. programs Mazzucato discusses in her book, such as DARPA and SBIR, are "dual use" programs: they support private sector research only in areas that are already in the research purview of federal mission agencies.

290. See *id.*; see also Lewis M. Branscomb, *The National Technology Policy Debate*, in *EMPOWERING TECHNOLOGY: IMPLEMENTING A U.S. STRATEGY* 8, 8–9, 14, 19 (Lewis M. Branscomb ed., 1993) (discussing objections to so-called "industrial policy" and "picking winners" in the context of U.S. technology policy).

291. Maryann P. Feldman & Maryellen Kelley, *Leveraging Research and Development: Assessing the Impact of the U.S. Advanced Technology Program*, 20 SMALL BUS. ECON. 153, 162 (2003); see also WENDY SCHACHT, CONG. RESEARCH SERV., 95-36, *THE ADVANCED TECHNOLOGY PROGRAM* (2007).

292. See WENDY H. SCHACHT, CONG. RESEARCH SERV., 95-30, *THE NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY: AN APPROPRIATIONS OVERVIEW* (2013) (summarizing the appropriations history of the ATP program). Note that in 2007 the name changed from the Advanced Technology Program to the Technology Innovation Program and some of the program's eligibility requirements shifted. See *id.* at 2.

criticized the federal government after it provided public financing to Solyndra, a solar company that failed.²⁹³

Lewis Branscomb, former director of NIST, describes this landscape as the U.S. government's "mission-justified" approach to technology policy. Under the mission-justified approach, the government avoids intervening in the private sector except where justified by a national mission and circumscribed by the goals and administrative capacity of a federal mission agency.²⁹⁴ In other words, Branscomb interprets the mission-justified approach as a pro-market, anti-interventionist policy.

The internalization principle—that public goods should be supplied by the smallest jurisdiction that internalizes the benefits—provides a different explanation for the federal government's selective, mission-oriented innovation policy. It's not just about adherence to the free market; it's about adherence to federalism and the role of states and local governments in controlling innovation policy. As explained in Section II.C, innovation, and particularly innovation that engages the private sector, tends to result in significant localized benefits for the region in which it occurs. Thus, while the federal government can justify using national money to finance R&D in "national public goods" areas that provide countrywide benefits, justifying federal funding for research outside of these national mission areas is much harder. First, it may be unfair for residents of, say, North Carolina to fund commercial software research that California companies are mostly conducting.²⁹⁵ Second, the efficiency arguments discussed in Section IV.C suggest that the national government should have stronger incentives and better information than local governments to fund research in areas where national agencies are already engaged. Absent federal funding, it is unlikely that states would fund research into, say, defense or public health of their own accord.

However, for general private sector innovation, the case is different. Here, the incentives are reversed: Congress's incentives to promote private sector innovation are highly diffuse, as economic benefits are spread about the country and highly concentrated in certain areas in the short term; but local governments' incentives to spend on private sector innovation are quite strong. This result is what one would expect from the collective action

293. Joe Stephens & Carol D. Leonnig, *Solyndra: Politics Infused Obama Energy Programs*, WASH. POST (Dec. 25, 2011), http://www.washingtonpost.com/solyndra-politics-infused-obama-energy-programs/2011/12/14/gIQA4HIIHP_story.html.

294. See Branscomb, *supra* note 290, at 14, 19; see also BRANSCOMB & AUERSWALD, *supra* note 28, at 144 ("[F]ederal politics views with suspicion government programs to assist individual firms.").

295. As mentioned, however, this is a controversial argument. See discussion *supra* Section III.C.

federalism theory: the smallest jurisdiction that captures the benefits is the most likely to act to obtain them. While states cannot effectively protect exclusive rights in public information that easily spills over to other jurisdictions, they can—and do—use public finance to internalize many of the benefits of innovation. States accomplish this internalization through direct taxation. They also internalize the variety of benefits that result from having innovators located in the region: more skills and knowledge within the populace, more employment and business opportunities, more start-ups, and more commercialization and production activities. The next section discusses this phenomenon in detail.

2. *U.S. State Innovation Finance*

The federal government’s “mission-oriented” approach to innovation finance stands in contrast to that of subnational governments such as states. As Branscomb and Philip Auerswald have observed, unlike the federal government, the states “are politically quite comfortable competing with one another to attract new business through active programs of R&D subsidies.”²⁹⁶ In this section, I assess these competing state R&D programs. Looking at programs from a wide selection of states, I show that, unlike federal innovation finance, state innovation finance can be available at all phases of development, including the commercialization stage, and is not limited to discrete technology areas—so long as the subject matter meets states’ economic development objectives.

a) *Origins*

State governments’ active efforts to promote innovation in their jurisdictions have deep roots in state economic development policy. The colonies and independent states granted “exclusive privileges” (the early term for patents and other exclusive rights) and a variety of other incentives, such as tax breaks and land grants, for companies and individuals to undertake high-cost development projects, such as iron mines and mills.²⁹⁷ Following ratification of the Constitution, states continued to use public debt to stimulate growth within their borders. States built roads, canals and transportation systems, and used a variety of tax inducements to spur regional economic growth.²⁹⁸ Indeed, throughout the nineteenth century,

296. BRANSCOMB & AUERSWALD, *supra* note 28, at 144.

297. On state and colonial patent practices, see Hrdy, *State Patent Laws in the Age of Laissez Faire*, *supra* note 25.

298. MICHAEL LIND, *LAND OF PROMISE* 53 (2013) (“In 1817, [Governor] Clinton persuaded the New York legislature to authorize \$7 million to build the canal. Half of the bonds for the Erie Canal were purchased by foreigners.”).

states, not the federal government, were the main drivers of economic development, using raised or borrowed money to finance transportation systems like canals and railroads.²⁹⁹ Under the Morrill Act of 1862, Congress allocated the states land on which to establish colleges and “agricultural experiment stations” to conduct research adapted to local crops and environmental conditions.³⁰⁰ Thereafter, states used their influence over universities to push applied research with practical implications for local economic development.³⁰¹

In the past thirty years, states have increasingly emphasized science and technology policy and created a variety of programs to encourage academic research and research-intensive enterprise in their jurisdictions.³⁰² According to Dan Berglund and Chris Coburn, whose 1994 study of state cooperative technology programs was one of the first to bring this phenomenon to light, the main reason for states’ growing interest was the increasing economic importance of knowledge-intensive industries as compared to manufacturing. By expanding their economic development policies to include support for research universities, corporate R&D, patent generators, and winners of federal grants, states sought to avoid the job and revenue losses that occur when manufacturing companies go bankrupt or leave.³⁰³

To achieve this end, several states in the 1980s established cooperative technology programs: “public-private initiatives involving government and industry—and often universities—that sponsor the development and use of technology and improved practices to measurably benefit specific

299. See generally Harry N. Scheiber, *State Law and “Industrial Policy” in American Development, 1790–1987*, 75 CALIF. L. REV. 415 (1987).

300. Under the Land-Grant Collect Act of 1862 (also known as the “Morrill Act”), Congress allocated federal land to each state to support development of a colleges focused on instruction of “agriculture and the mechanic arts.” See 7 U.S.C. § 301 et seq. (2012). Under the 1887 Hatch Act, more land was given to the states for Agricultural Experiment Stations associated with land grant colleges. See 7 U.S.C. § 361a et seq. (2012); Tiffany Shih & Brian Wright, *Agricultural Innovation*, in ACCELERATING ENERGY INNOVATION: INSIGHTS FROM MULTIPLE SECTORS 55-57 (Rebecca M. Henderson & Richard G. Newell eds., 2011) (discussing the establishment of land grant colleges and state agricultural experiment stations in the United States); see also Fromer, *The Intellectual Property Clause’s External Limitations*, *supra* note 25, at 1348–49, 1356 (noting that one of the reasons the Framers decided not to give Congress express authority power to establish universities was the perception that states could do so on their own).

301. See David Audstretch, *The Entrepreneurial Society and the Role of the University*, 32 ECONOMIA MARCHE J. APPLIED ECON. 7, 8 (2013); Peter Lee, *Patents and the University*, 63 DUKE L.J. 1, 8–10 (2013).

302. BERGLUND & COBURN, *supra* note 201, at 5–9.

303. *Id.*

companies” for the goal of economic growth.³⁰⁴ Examples include Pennsylvania’s Ben Franklin Partnership and Oklahoma’s Center for Advancement of Science and Technology.³⁰⁵ By the early 1990s, *all fifty states had such programs*.³⁰⁶

b) An Emphasis on Localized Benefits

After reviewing all the programs available as of 1994, Berglund and Coburn came to the conclusion that state innovation programs are fundamentally different from federal innovation programs because they seek *localized* benefits. A “key criterion” for states, they write, is “the degree to which the projected benefits can be captured in the target region.”³⁰⁷ Specifically, as others have noted, states hope to obtain “gains in employment; diversification of the regional economy; the influx and retention of a highly educated labor force; an expansion of the tax base; and growth in related service industries.”³⁰⁸ So “[w]hile state-sponsored programs may have benefits beyond their borders, states are aggressive about ensuring that they capture an appropriate share within their borders,” which is “reflected in the types and stages of research and technology investments states make.”³⁰⁹ As a result, “[t]his orientation drives states toward industrially related technology fields where new products and processes can be readily deployed by companies resident in the state.”³¹⁰

For example, when Connecticut established Connecticut Innovations in 1989³¹¹ to invest in local start-ups, the legislature explicitly based its decision to spend such large sums of public money on the finding “that the creation of new technology-based businesses represents an important source

304. *Id.* at 1.

305. *Id.* at 8–9.

306. *Id.* at 9.

307. *Id.* at 15.

308. Terrance McGuire, Note, *A Blueprint for Disaster? State Sponsored Venture Capital Funds for High Technology Ventures*, HARV. J.L. & TECH. 419, 419 (1994); see also Bo Zhao & Rosemarie Ham Ziedonis, *State Governments as Financiers of Technology Startups: Implications for Firm Performance*, at 5 (July 2012) (unpublished manuscript), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2060739 (observing that in financing innovation “state governments pursue more parochial interests: to stimulate economic growth inside geographic borders and to diversify the tax base”); BERGLUND & COBURN, *supra* note 201, at 5–9 (discussing the origin of state technology initiatives and noting that these initiatives were closely linked to states’ long-standing interest in promoting economic development).

309. BERGLUND & COBURN, *supra* note 201, at 15.

310. *Id.*

311. See “Innovation Capital Act of 1989,” codified at CONN. GEN. STAT. ANN. § 32-32 et seq. (West 2010 & Supp. 2014).

of new jobs for the economy of the state, that it is essential for existing businesses and industry to innovate and adopt new state-of-the-art processes and technologies in order for such business and industry to expand, to create and retain employment and to better compete in the global marketplace.”³¹²

On the flip side, states are unlikely to support non-commercial innovations or basic research with little current commercial relevance “because its results are so easily diffused before any special benefit can be gained.”³¹³ Instead, states would be more likely to focus on research with near-term commercial relevance for local companies.³¹⁴ States hope any benefits resulting from this research will “stick to the ribs” of local firms for as long as possible rather than spilling over to other firms outside the state.³¹⁵ In addition to the threat of knowledge spillover, states are also motivated by the threat of “brain-drain”: the departure of the region’s most highly skilled people.³¹⁶

c) Categories of State Innovation Finance

Today all states and many cities make promoting innovation a core feature of their economic development policies. Local innovation strategies vary widely and are constantly evolving. Five categories of state innovation finance exist: (1) research incentives, (2) commercialization incentives, (3) R&D tax incentives and subsidies, (4) education and worker training programs, and (5) investments in infrastructure. Each state finance category has a similar two-fold goal: to promote investment in innovation, i.e. doing things that are new to some unit of adoption and that have “value” in some

312. CONN. GEN. STAT. ANN. § 32-33. The state was particularly concerned about Connecticut’s overreliance on defense contracts with the federal government. *See id.* (“It is further found and declared that Connecticut ranks very high among the states on a per capita basis in the amount of prime defense contracts awarded; that the economies of many areas in the state and the employment opportunities offered by many businesses in the state are heavily defense-dependent and would suffer severe adverse impacts in the event of prime defense contract cutbacks . . .”).

313. BERGLUND & COBURN, *supra* note 201, at 15.

314. *See id.* Of course, as Berglund and Coburn note, the federal government also seeks to capture benefits from its investments in the United States. This “federal approach is reflected in program selection criteria.” *Id.* For instance, federal selection criteria may include the requirement that research have relevance for a national mission such as defense. This is called a “dual-use” technology, as explained above. *See supra* Section V.B.1.c.

315. *See* Louis G. Tornatsky, *Building State Economies by Promoting University-Industry Technology Transfer*, at 10 (2000), <http://www.gbcbiotech.com/transfereciatecnologia/assets/building-state-economies-by-promoting-university-industry-technology-transfer.pdf> (published by the Nat’l Governors Ass’n Ctr. for Best Practices).

316. *See id.* at 10 (“There is evidence of a considerable imbalance in the across states in the interstate migration of highly skilled people, commonly referred to as the ‘brain drain.’”).

commercial market,³¹⁷ and to achieve the localized benefits just mentioned: employment opportunities for residents, attracting a skilled workforce, diversification of the economy and the tax base, and growth in related industries like construction, specialized suppliers, and restaurants.

i) Research Incentives

As discussed, basic research funding is mostly federal and states are unlikely to spend large amounts of public money on basic research with little near-term commercial relevance. But states have nonetheless taken on an increasing share of the responsibility for funding science and technology-based research. Based on years of research studying states' growing engagement in science and technology policy, Maryann Feldman has concluded that the states' growing role in the "basic research enterprise" suggests that "public support for R&D no longer rests solely at the federal level."³¹⁸

In recent research, Feldman and Lauren Lanahan report that since 1980 states' expenditures on university R&D programs have increased threefold to \$3.13 billion, which accounts for 5.8% of all university research and is more funding than industry supplies for academic R&D.³¹⁹ Building on Berglund and Coburn's 1994 survey of state technology programs, Feldman and Lanahan document several trends in "state science policy experiments."³²⁰

They identify three types of common state programs that support basic and applied research at universities. Each of these programs appear to be explained by states' interests in capturing the benefits of research and commercialization within their borders.³²¹ In addition, the chance for capturing federal grant money is also a driving factor. Maybe this is

317. See definition of innovation *supra* Section II.A.

318. Feldman & Lanahan, *supra* note 29, at 287. For empirical work on state financing for innovation, see, for example, Maryann Feldman & Lauren Lanahan, *Silos of Small Beer – A Case Study of the Efficacy of Federal Innovation Programs in a Key Midwest Regional Economy*, CTR. FOR AM. PROGRESS, at 3–4 (Sept. 2010), https://www.americanprogress.org/wp-content/uploads/issues/2010/09/pdf/small_beer_exec_summary.pdf (finding that state awards were perceived to be more accessible than federal awards); Zhao & Ziedonis, *supra* note 308, at 2–3 (concluding Michigan's technology financing grants enhanced company survival as compared to similar prospects that did not get award).

319. Feldman & Lanahan, *supra* note 29, at 287.

320. *Id.* at 288.

321. *Id.* at 287–88 (asserting that states invest in science as a means to "facilitate commercialization" and "capture returns within their borders" and to "increase their share of federal R&D expenditures").

cooperative federalism in action: states are paying for the local benefits, and the federal government is picking up the bill for out-of-state spillovers.³²²

1. Recruitment Incentives for Faculty and “Star Scientists.” The first program, which Feldman and Lanahan call eminent scholars programs, creates recruitment incentives for prolific “star scientists” with strong records in obtaining grants and patents. Grants are usually \$3-6 million per scholar, and as of 2009, 21 states have adopted eminent scholars programs.³²³ For example, Arizona has an “eminent scholars matching grant fund” that allocates some of the state’s annual income towards matching funds “to attract and retain eminent faculty.” The funds provide a certain amount of matching for “nonpublic endowment monies donated [to the universities] to attract and retain eminent faculty.”³²⁴ The idea behind these programs is to invest in the people who generate basic research and who win federal funding and patents, rather than the research itself. In other words, they have the goal of producing human capital externalities, to use Moretti’s term, rather than simply producing new technical knowledge.

2. Grants for Research. The second common state incentive for academic research, adopted in 29 states as of 2009, is university research grants programs. These programs create state-sponsored grants for basic and/or applied research, available to all researchers at universities or research institutions in the state, without requiring prior federal funding or other non-state matching.³²⁵ According to Berglund and Coburn, who noted the emergence of state basic and applied research grants programs in the 1980s, the purpose of state basic research grants is to help scientists who were “on the verge of becoming nationally competitive in receiving funding” by providing them grants early in their careers. Additionally, the state basic research grants create “a track record that will help them to compete for federal monies, thereby bringing more research funds to the state.”³²⁶ Grants for applied research, meanwhile, are “deigned to help scientists become active in applied research and to encourage partnerships

322. See discussion of matching grants to compensate for externalities *supra* Section IV.D.1.

323. Feldman & Lanahan, *supra* note 29, at 290.

324. ARIZ. REV. STAT. § 15-1663.B (LexisNexis 2015).

325. See, e.g., *id.*

326. BERGLUND & COBURN, *supra* note 201, at 84 (discussed in Feldman & Lanahan, *supra* note 29, at 291).

with industry, which, in turn, will benefit from the new technology that creates and retains jobs.”³²⁷

3. Centers to Increase University-Industry Cooperation. The third program identified by Feldman and Lanahan is the “Center for Excellence”—university-based centers with a focus on promoting collaboration between universities and industry and on encouraging faculty to undertake research oriented towards the needs of specific industries or technologies.³²⁸ For instance, Massachusetts created the Massachusetts Centers of Excellence Corporation (now the Biotechnology Center of Excellence Corporation (BCEC)) in 1985 to “facilitate technology transfer and commercialization of emerging technologies through university/industry collaboration.”³²⁹ According to Feldman and Lanahan, this type of program is the most common, adopted by 28 states, and also the first type of program that states adopted.³³⁰ Their explanation is that, generally speaking, the states are more comfortable “[prioritizing] making investments in academic research *directly linked to industrial activity over supporting more upstream efforts.*”³³¹

ii) Local Commercialization Incentives

The theme that emerges from state investments in basic and academic research is that states are driven not only by the desire to correct market failures in research, but also by the desire to “stimulate economic growth inside geographic borders.”³³² Thus, it should not be surprising that state actions in innovation finance become far more visible, and far more expensive, when we look at incentives whose express purpose is to resolve market failures in *commercialization* of inventions, rather simply to generate new knowledge.³³³ This section describes several kinds of commercialization incentives offered by states and sometimes cities. The main division between local commercialization incentives is whether they

327. *Id.*; see also Ark. Econ. Dev. Comm. Science & Tech., About Commercialization: Technology Transfer Assistance Grant Program (TTAG), <http://www.asta.arkansas.gov/ttag.html> (2015).

328. Feldman & Lanahan, *supra* note 29, at 292.

329. See BIOTECH. CTR. EXCELLENCE CORP., <http://home.mindspring.com/~bcec/> (last visited Feb. 9, 2016).

330. Feldman & Lanahan, *supra* note 29, at 292.

331. *Id.* (emphasis added).

332. See Zhao & Ziedonis, *supra* note 308, at 5 (discussing various reasons Michigan likely began supplying financing for local technology companies, including both correcting market failures and improving the regional economy).

333. See Hrdy, *Commercialization Awards*, *supra* note 5, at 21.

are directed at the university interface or at purely private enterprises such as new start-ups seeking financing.

1. Technology Transfer Offices. States' involvement in technology transfer at the university interface illustrates concerted effort to push academic research to market: the transformation of academic research into commercial products, services, or other applications.³³⁴ Universities usually have technology transfer offices (TTOs) that assist university faculty in patenting their inventions and then sell or license those patents to private sector firms. Patent and other IP often (though not necessarily) mediate technology transfer based on the theory that, if the underlying research can be protected, this protection will make private companies more likely to invest in its commercialization.³³⁵ Ostensibly, TTOs aim to ensure that university-generated research is put into use “for the broader benefit of society.” But in practice, TTOs may be more focused on generating revenues for the university than on ensuring university research is commercialized or creates tangible benefits for the community.³³⁶

Besides the university and society at large, technology transfer offices can benefit the local economy as a result of the licensing activity that occurs in and around universities and government labs. Not only can local governments collect taxes on profits, sales, payrolls, and property from universities and licensee corporations, they can also expect benefits from spillover to local businesses in non-innovation sectors, especially through employment.³³⁷ For instance, in 2014, the University of California system reportedly obtained 496 patents, disclosed 1796 inventions, and produced 86 start-ups (faculty “spin-outs”). These in turn generated \$14 billion in revenue for the university (most of which goes back into university

334. Arti Rai & Rebecca Eisenberg, *Bayh-Dole Reform and the Progress of Biomedicine*, 66 L. & CONTEMPORARY PROBLEMS 289 (2003).

335. *But see* Eisenberg, *Limiting the Role of Patents in Technology Transfer*, 37 L. QUAD. NOTES 40, 40–43 (1994) (adapted from remarks presented to the Congressional Biomedical Research Caucus in Washington, D.C., on June 28, 1993) (questioning the wisdom of allowing patents on government-sponsored research to promote technology transfer and suggesting that costs of patenting government research may outweigh the benefits).

336. *See* Jay Kesan, *Transferring Innovation*, 77 FORDHAM L. REV. 2169, 2169 (2009) (arguing that university technology transfer activities are “predominantly patent-centric and revenue-driven with a single-minded focus on generating licensing income and obtaining reimbursement for legal expenses”); *see also* Liza Vertinsky, 4 UTAH L. REV. 1949, 1949 (2012) (arguing that “if universities obtained more discretion, responsibility, and accountability over the post-discovery development paths for their inventions, they might be able to improve the trajectory for many promising scientific discoveries”).

337. MORETTI, *supra* note 40, at 73–120 (discussing the variety of tax and labor benefits that accrue in those regions that succeed in developing innovation clusters).

operations) and led to the employment of around 19,000 people in the area.³³⁸ Many doubt whether universities generate enough revenues to cover their costs in technology transfer.³³⁹ But whether the benefits to the local economy justify the costs is more difficult to measure or to dispute.³⁴⁰

2. Venture Financing – University Spin Outs. Along with TTOs, some universities also provide venture financing to university faculty and students seeking starting their own companies outside the university, called spin-outs. For example, the University of California is preparing to launch “UC Ventures,” an independent \$250 million fund to pursue investments in enterprises fueled by UC research. According to proponents like UC President Janet Napolitano, UC Ventures will spur technology commercialization efforts already underway at UC schools and will help faculty and students “develop innovations that can benefit California and the world.”³⁴¹

3. Venture Financing – Private Enterprises. State commercialization initiatives are not limited to the university interface. Recently states, and even some cities, have begun funding their own venture capital units to supply *private enterprises* with financing for commercializing inventions and testing out new business models.³⁴² State venture capital has two purposes: generating revenues and boosting local economic development through spillovers. When states obtain equity stakes in the companies they fund, they can theoretically achieve a good return on investment and generate profits for the state.³⁴³ However, as Terrance P. McGuire has observed, this goal of achieving a high return on the investment may conflict

338. UC Office of the President Press Release, Technology Commercialization Report (FY 2014), at 16, <http://www.ucop.edu/innovation-alliances-services/innovation/innovation-impact/technology-commercialization-report.html>.

339. See, e.g., Walter Valdivia, *University Start-Ups: Critical for Improving Technology Transfer*, BROOKINGS INST. CTR. FOR TECH. INNOVATION, at 1 (Nov. 2013), <https://www.brookings.edu/research/university-start-ups-critical-for-improving-technology-transfer/>.

340. See, e.g., Maryann P. Feldman & Pierre Desrochers, *Research Universities and Local Economic Development: Lessons from the History of the Johns Hopkins University*, 10 INDUS. & INNOVATION 5 (2003).

341. UC Office of the President, University of California Proposes Creation of New Venture Fund to Invest in UC Innovation, UNIV. OF CAL. (Sept. 15, 2014), <http://www.universityofcalifornia.edu/press-room/university-california-proposes-creation-new-venture-fund-invest-uc-innovation>.

342. Hrdy, *Commercialization Awards*, *supra* note 5, at 54–56.

343. See Terrance McGuire, *supra* note 308, at 435.

with the goal of promoting local economic development.³⁴⁴ For example, if a state decides to focus on the highest-profit investments, it might fund companies that hire few local workers or that do not intend to remain in the state for a long period. Alternatively, a state more concerned about promoting local development should fund a company with a solid base in the local economy that is likely to remain, hire local employees, and generate positive externalities for other companies in the area that benefit from the same suppliers and labor markets.

iii) Business Tax Incentives and Subsidies

As is now clear, states' interest in science and technology policy extends outside the walls of the university. The best examples are state subsidies and tax incentives for research firms that agree to locate in the state. In pursuit of the economic benefits of research-intensive industries, states and cities spend billions of dollars every year on tax incentives and subsidies to attract businesses, usually qualifying such incentives on the requirement that recipients perform qualifying research and development in the state.³⁴⁵ As Walter Hellerstein and Dan Coenen have discussed, these subsidies serve as an "inducement to local industrial development and expansion."³⁴⁶

As of 2010, 34 states offer general R&D tax credits.³⁴⁷ Most state R&D tax credits mirror all or some aspects of the federal R&D credit.³⁴⁸ Under the federal R&D tax credit, contained in Section 41 of the U.S. Internal Revenue Code, businesses can obtain a tax credit based on qualifying research expenses so long as they are (1) undertaken for purposes that are "technological in nature," (2) intended to yield applications "useful in the development of a new or improved business component," and (3) comprise activities "substantially all . . . of which constitute elements of a process of

344. *Id.* ("[T]he fundamental question that confronts all state planners [starting a public venture fund] is whether the fund should focus on return on investment (ROI) or economically targeted investments (ETI).").

345. For a survey of state R&D tax incentives, see MICHAEL D. RASHKIN, RESEARCH AND DEVELOPMENT TAX INCENTIVES: FEDERAL, STATE, AND FOREIGN 1–2 (2007). *See also* Legislative Budget Board, State of Texas, Overview of Research and Development Tax Incentives 14–23 (2013) (surveying business ad sales tax incentives in all fifty states).

346. Walter Hellerstein & Dan Coenen, *Commerce Clause Restraints on State Business Development Incentives*, 81 CORNELL L. REV. 789, 790 (1996).

347. For a state-by-state survey, see RASHKIN, *supra* note 345. *See also* Moretti & Wilson, *supra* note 87, at 3.

348. *See* RASHKIN, *supra* note 345, at 265 (noting that most state research credits are based on the federal credit albeit with significant variations from state to state); *see also* Ruth Mason, *Delegating Up: State Conformity with the Federal Tax Base*, 62 DUKE L.J. 1267, 1274–79 (discussing states' tendency to conform their tax laws to the federal government's).

experimentation.”³⁴⁹ Along with offering a state-level version of this basic R&D credit, some states provide additional tax incentives that are not currently available at the federal level, such as refundable credits for start-ups and early-stage companies that have no profits yet and cannot benefit from the federal credit.³⁵⁰ State start-up credits have spurred calls for a similar federal credit.³⁵¹

iv) Education and Worker Training

The largest state innovation finance expenditure is state funding for education.³⁵² In recent years, states have become increasingly motivated to support Science, Technology, Engineering, and Math (STEM) education. As Jonathan Rothwell writes in a recent Brookings Institute report on STEM education, “state and local governments affect STEM education through many channels. They boost university and community college STEM education through funding and scholarships. They support training by coordinating workforce development efforts, and they shape K-12 STEM education by approving and funding of STEM-focused schools; the training, certification, and management of teachers; and the development and enforcement of content standards.”³⁵³

349. I.R.C. § 41(d)(1) (2012).

350. Several states offer refundable R&D tax credits. For example, Louisiana offers an R&D tax credit that can be carried forward for up to five years. LA. STAT. ANN. § 47:6015(K) (West 2016) (“If the amount of the [R&D credit authorized under this Section] exceeds the amount of tax liability for the tax year, the excess credit may be carried forward as a credit against subsequent Louisiana income or corporation franchise tax liability for a period not to exceed five years.”). For more examples of refundable state R&D tax incentives, see Joe Stoddard, *States Battle for R&D Investment by Enhancing Tax Incentives*, THE TAX ADVISER (Jan. 31, 2012), <http://www.thetaxadviser.com/issues/2012/feb/clinic-story-10.html>.

351. See Sen. Chris Coons & Sen. Mike Enzi, *R&D Tax Credit Spurs Innovation*, POLITICO (Mar. 7, 2013), <http://www.politico.com/story/2013/03/rd-tax-credit-spurs-innovation-088525> (calling for support for the Startup Innovation Credit Act, S. 193, 113th Cong. (2013)).

352. In 2012, local education agencies in the 50 states and the District of Columbia reported \$603.5 billion in total revenues on education. Of those revenues, only \$60.7 billion (10.1%) came from the federal government. \$272.4 billion (45.1%) came from state governments, and \$270.4 billion (44.8%) came from local government. U.S. DEP’T OF EDUC.: NAT’L CTR. EDUC. STATISTICS, REVENUES AND EXPENDITURES FOR PUBLIC ELEMENTARY AND SECONDARY SCHOOL DISTRICTS: SCHOOL YEAR 2011–12 (FISCAL YEAR 2012), at 4, tbl. 1, “Sources of revenues and type of expenditures for public elementary and secondary education, by state or jurisdiction: Fiscal year 2012,” <https://nces.ed.gov/pubs2014/2014301.pdf>.

353. Jonathan Rothwell, *The Hidden STEM Economy*, BROOKINGS INST. METRO. POL’Y PROGRAM (June 2013), at 20–21, <http://www.brookings.edu/~media/research/files/reports/2013/06/10-stem-economy-rothwell/thehiddenstemeconomy610.pdf>; see also

Examples of state programs focused on STEM education include: Illinois' "STEM Learning Exchanges," which establish contracts for regional, educational and business networks to assess educational needs and confront challenges for STEM education;³⁵⁴ North Carolina's "Education Enhancement Grants" for non-profit institutions that develop programs to enhance biotechnology education and workforce training;³⁵⁵ and Georgia's Educational Technology Consortium, which operates several programs to improve students' access to technology and development of skills useful in high tech jobs.³⁵⁶

Why are states funding these programs? Economic federalism theory suggests an explanation. Perhaps, like knowledge itself, education in knowledge-intensive fields behaves like a "local public good": a resource that primarily benefits a particular geographic region, and that may be best supplied at a "very local level."³⁵⁷ Even though STEM education can obviously produce significant national spillovers when people share knowledge across state borders or move to work in other states, the beneficiaries of these programs may well decide to reside in the region and become employed by local firms. In other words, local policymakers have a strong incentive to supply STEM education because they believe they can internalize the benefits. At the same time, the pressure to compete with other states creates an additional incentive to support STEM education: to avoid flight.

STATE SCI. & TECH. INST., TRENDS IN TECHNOLOGY-BASED ECONOMIC DEVELOPMENT: LOCAL, STATE AND FEDERAL ACTION IN 2012, at 10–12 (2012), <http://ssti.org/sites/default/files/trends2012.pdf> (discussing recent STEM initiatives offered by states and cities); Aaron Chatterji, Edward Glaeser, & William Kerr, *Clusters of Entrepreneurship and Innovation*, 14 INNOVATION POL'Y & ECON. 129, 152 (2014) ("At the state level, 18 states have passed some legislation as of 2007 to encourage entrepreneurship education, with significant variance in terms of requirements and curriculum.").

354. See 105 ILL. COMP. STAT. 5/26-17(3) (2015); ILL. PATHWAYS, *STEM Learning Exchanges*, <https://www.illinoisworknet.com/ilpathways/Pages/STEMLE.aspx> (last visited Feb. 2, 2017).

355. See N.C. BIOTECH. CTR., EDUCATION ENHANCEMENT GRANT PROGRAM, <http://www.ncbiotech.org/workforce-education/education-funding/education-enhancement-grant> (last visited Feb. 9, 2016); see also N.C. BIOTECH. CTR., EDUCATION ENHANCEMENT GRANTS: FULL PROPOSAL GUIDELINES AND INSTRUCTIONS (June 21, 2013), <http://www.ncbiotech.org/sites/default/files/funding/2013-2014%20EEG%20FP%20Guidelines%20and%20InstructionsFinal.pdf>.

356. GA. EDUC. TECH. CONSORTIUM, INC., INNOVATION GRANT WINNERS 2014, <http://www.gaetc.org/domain/133> (last visited Feb. 9, 2016).

357. See MUELLER, *supra* note 151, at 81 (stating that schooling is a public good that is typically or at least feasibly could be provided "at a very local level").

v) Infrastructure and Public Services

States and cities finance the lion's share of physical infrastructure and public services, such as bridges, roads, and parks. These physical infrastructure and public services are classic examples of a "local public good" that suffers from congestion and is more efficiently provisioned at the local level.³⁵⁸ In undeveloped countries, basic infrastructure is an obvious component of any policy for promoting economic development and innovation.³⁵⁹ In the United States, where basic infrastructure is a given, states and cities can still influence levels of innovation in the region by tailoring their infrastructure to entrepreneurs and high-innovation sectors.

States have long supported university facilities like lab space.³⁶⁰ Another example of physical infrastructure is the science or research "park": a formally planned cluster of innovative businesses and institutions, typically centered around one or more universities or national labs.³⁶¹ Famous research parks include Research Triangle Park in Raleigh-Durham, North Carolina, the largest research park in the world.³⁶² Research parks also come in smaller sizes and are called different names without changing the fundamental idea behind them: to bring companies and researchers together in one place. For instance, "as part of its long-term economic development focus," Davis, CA is building several new "innovation centers": "clusters of technology companies located together to create a kind of critical mass for new ideas and new products."³⁶³ The city hopes that, although the centers may take "years to build out," it will "eventually

358. In 2007, local governments spent \$145 billion (6% of revenues) on highways, new roads, and maintenance, and \$955 billion on other public services, such as libraries, police, garbage removal, fire protection, park maintenance, snow removal, etc. EXEC. OFFICE OF THE PRESIDENT, ECONOMIC REPORT OF THE PRESIDENT, 2010-431 (2010), tbl. B-86, "State and local government revenues and expenditures, selected fiscal years, 1942-2007"; *see also* MUELLER, *supra* note 151, at 81-82.

359. *See generally* ROBERT D. COOTER & HANS-BERND SCHÄFER, SOLOMON'S KNOT: HOW LAW CAN END THE POVERTY OF NATIONS (2012).

360. NAT'L SCI. BD., SCIENCE AND ENGINEERING INDICATORS 2014, CHAPTER FIVE: ACADEMIC RESEARCH AND DEVELOPMENT 10-14 (2012) (reporting that state and local governments spend around \$3 billion on university facilities and research and various R&D programs, which is more than is provided by industry).

361. BRUCE KATZ & JENNIFER BRADLEY, THE METROPOLITAN REVOLUTION 113-14 (2013) (discussing the prominent role of cities in cluster strategies).

362. *See* RESEARCH TRIANGLE PARK, WHO WE ARE, <http://www.rtp.org/about-us/> (last visited Feb. 9, 2017).

363. Dave Ryan, *City Hires New Chief Innovation Officer*, THE DAVIS ENTER., May 31, 2015, at A1.

bring in large amounts of property and sales taxes, as well as high-paying jobs.”³⁶⁴

Outside the university, cities provide public services that can make or break the successful development of a new technology or innovative business plan. A high-profile example is city provisioning of broadband and high-speed Internet to encourage entrepreneurship in the region.³⁶⁵ Dozens of cities across the country are setting up municipal broadband networks.³⁶⁶ Since broadband, lab space, and research parks all involve a physical location, it is not controversial to suggest local governments should at least partly finance them.³⁶⁷ To the extent that these programs create national knowledge spillovers, the federal government should match funds. Broadband is an example of a technology that, despite its linkage to local infrastructure, could easily lead to uncontrollable knowledge spillovers and warrant federal sponsorship. For example, the White House’s new initiative, BroadbandUSA, operated through the Department of Commerce, will promote broadband deployment and adoption in undersupplied cities around the country using federal funds allocated in the American Recovery and Reinvestment Act of 2009.³⁶⁸

VI. CONCLUSION

In light of widespread dissatisfaction with IP rights, many claim innovation finance should supplant IP. The federal government, they argue, should supply this financing. But this conflicts with what currently happens. In the United States, patent law is federal; yet outside select research areas—primarily defense—funding for private sector innovation frequently comes from states and other subnational governments. For better or worse, U.S. innovation policy operates at both the federal and state level. This situation can be explained, and arguably justified, based on principles of economic federalism, under which innovation finance should be supplied

364. *Id.*

365. See Susan Crawford, *How Cities Can Take on Big Cable*, BLOOMBERG VIEW (June 27, 2014), <http://www.bloombergview.com/articles/2014-06-27/how-cities-can-take-on-big-cable>; Susan Crawford, *The Wire Next Time*, N.Y. TIMES, Apr. 28, 2014, at A21.

366. See Steve Lohr, *Lack of Choice Led to Push for Net Neutrality*, N.Y. TIMES, Feb. 26, 2015, at B1, B4.

367. See COOTER, *supra* note 16, at 105; see also Olivier Sylvain, *Broadband Localism*, 73 OHIO ST. L.J. 795 (2012) (“[L]ocal governments are supplying broadband service to residents to fill the service gap left by major providers.”).

368. Press Release, The White House, FACT SHEET: Broadband that Works: Promoting Competition & Local Choice in Next-Generation Connectivity (Jan. 13, 2015), <https://www.whitehouse.gov/the-press-office/2015/01/13/fact-sheet-broadband-works-promoting-competition-local-choice-next-gener>.

by the smallest level of government that internalizes the benefits of its efforts. While subnational governments cannot realistically internalize the benefits of patent regimes that result in widespread diffusion of new information, they can internalize meaningful benefits from innovation by strategically financing it.

This conclusion has several major implications. First, if direct financing for innovation is chosen in favor of patent rights, a powerful beat of localism may be inevitable. Given that local governments can directly internalize many of the economic benefits of innovation finance, they are likely to have exceptionally strong incentives to subsidize private-sector innovations that are expected to benefit local firms and residents.

Second, economics—and specifically the economic theory of federalism—suggests that this division of authority represents a more optimal allocation of responsibilities between state and federal governments.³⁶⁹ *Precisely because* their residents are the ones who directly benefit, local governments should have better incentives, and also better information, to design and fund innovation incentives that will work for the region. In addition, inter-jurisdictional competition to attract mobile participants in innovation industries—high-tech firms, skilled talent, and related firms and institutions—should push local policymakers to craft policies that are more effective than their neighbors', and lead to more precise clustering of firms and talent to appropriate locations.³⁷⁰ Thus, if we really care about “growing innovation clusters for American prosperity,” we must care about state and local governments.³⁷¹

Lastly, the Article highlights that local governments' role in innovation finance has significant limits. The major limitation highlighted by the economic federalism literature is lingering externalities: obviously, national funding for knowledge with widespread social value is still necessary in cases where states cannot internalize sufficient economic benefits to justify funding it. In addition, as shown in Part II, geographic inequality among different regions in the U.S. may *already* be so severe that redistribution of benefits from rich to poor locations may be warranted—otherwise these regions will not realistically be able to compete in the “cluster competition” in the first place. Although economic federalism does not typically care about such arguments, I argue that geographic inequality too creates a strong argument for national intervention.³⁷²

369. *C.f.* Posner, *supra* note 144, at 41; *see also supra* Part IV.

370. *C.f.* COOTER, *supra* note 14, at 106, 129–30.

371. WESSNER, *supra* note 2.

372. *See* Hrdy, *Cluster Competition*, *supra* note 185.

USING ANTITRUST LAW TO CHALLENGE TURING'S DARAPRIM PRICE INCREASE

Michael A. Carrier,[†] Nicole L. Levidow^{††} & Aaron S. Kesselheim^{†††}

ABSTRACT

In 2015, notorious pharmaceutical entrepreneur Martin Shkreli made worldwide headlines. As CEO of Turing Pharmaceuticals, Shkreli increased the price of pyrimethamine (Daraprim) 5000 percent. Although Turing's price hike on the unpatented drug was met with widespread outrage, few recognized that the company had recently changed its distribution system from one in which the drug was widely available to one in which supplies could be obtained from only a single source. This Article contends that Turing's restricted distribution scheme for pyrimethamine, with its apparent lack of legitimate justifications, could form the basis of an antitrust violation. Turing appears to have monopoly power in engineering and maintaining a 5000 percent price increase, preventing hospitals from obtaining pyrimethamine, and ensuring the absence of FDA-approved substitutes for the drug. The company also appears to have engaged in exclusionary conduct when it changed its distribution system in a way that only made sense by blocking generic competition. Because the combination of monopoly power and exclusionary conduct is the hallmark of a monopolization claim, Turing's behavior warrants close antitrust scrutiny.

DOI: <https://dx.doi.org/10.15779/Z383R0PS73>

© 2016 Michael A. Carrier, Nicole L. Levidow & Aaron S. Kesselheim.

[†] Michael A. Carrier is a Distinguished Professor at Rutgers Law School.

^{††} Nicole L. Levidow is a Research Fellow with the Program On Regulation, Therapeutics, And Law (PORTAL) in the Department of Medicine, Division of Pharmacoepidemiology and Pharmacoeconomics, at Brigham and Women's Hospital.

^{†††} Aaron S. Kesselheim is an Associate Professor of Medicine at Harvard Medical School, and faculty in the Department of Medicine, Division of Pharmacoepidemiology and Pharmacoeconomics, at Brigham and Women's Hospital. Dr. Kesselheim's work is supported by the Laura and John Arnold Foundation, with additional support from the Engelberg Foundation and the Harvard Program in Therapeutic Science. The authors would like to thank Harry First, Herb Hovenkamp, Christopher Leslie, and Barak Richman for very helpful comments.

TABLE OF CONTENTS

I.	INTRODUCTION	1380
II.	GENERIC DRUG APPROVAL AND DISTRIBUTION SYSTEMS	1382
III.	MONOPOLY POWER	1386
	A. INDIRECT PROOF.....	1387
	B. DIRECT PROOF.....	1388
	C. MONOPOLY POWER NOT NEGATED BY COMPOUNDED VERSION	1389
IV.	EXCLUSIONARY CONDUCT	1393
	A. MONOPOLIZATION CASE LAW	1394
	B. RISK EVALUATION AND MITIGATION STRATEGY (REMS) CASE LAW	1395
	C. APPLICATION TO TURING.....	1398
	D. COUNTERARGUMENTS.....	1400
V.	ADDITIONAL EXAMPLES	1405
VI.	CONCLUSION	1408

I. INTRODUCTION

Notorious pharmaceutical entrepreneur Martin Shkreli made worldwide headlines in 2015. As CEO of Turing Pharmaceuticals, Shkreli obtained U.S. marketing rights to pyrimethamine (Daraprim) and quickly increased the price 5000 percent, from \$13.50 to \$750 per pill. Pyrimethamine is a decades-old drug used primarily to treat toxoplasmosis, a fatal parasitic brain infection that usually occurs in patients with weakened immune systems, such as those with end-stage HIV infection.

Turing's price hike was met with widespread outrage among the public and in the medical and public health communities, with the episode leading to censure by other drug companies, congressional hearings seeking ways to address the problem, and policy proposals from Republican and Democratic presidential candidates. Despite the fact that there were no patents or other forms of market exclusivity protecting the drug, Turing was able to raise the price because the relatively small market in the United

States for pyrimethamine had attracted no other generic manufacturers. Indeed, Shkreli later lamented that he did not raise the price even higher.¹

In addition to increasing price, Turing initiated another less widely understood move—it changed the distribution scheme for the drug. Before its acquisition by Turing, pyrimethamine was available without restriction to patients seeking to fill prescriptions at local pharmacies and to hospitals seeking to stock the product for inpatient use. But in the months before the price hike, apparently as a condition of the sale to Turing, pyrimethamine was switched to a controlled distribution system called Daraprim Direct, in which prescriptions or supplies of the product could be obtained only from a single source: Walgreen's Specialty Pharmacy.² As a result, hospitals could no longer obtain the drug from a general wholesaler, and patients could no longer find it at a local pharmacy. Instead, Turing required institutions and individuals to set up accounts through Daraprim Direct, and outpatients were only able to receive the drug by mail order.³ Comments from Turing executives suggest that a primary goal of the Daraprim Direct system was to make it impossible for anyone other than registered clients to obtain the drug, including generic manufacturers wishing to obtain samples for use in bioequivalence studies needed to obtain Food and Drug Administration (FDA) approval of their applications for generic versions.⁴

The central thesis of this Article's analysis is that Turing's restricted distribution scheme for pyrimethamine, with its apparent lack of legitimate justifications, could form the basis for an antitrust violation, especially if the scheme was established to prevent subsequent entrants into the market from undercutting the newly established high price for the drug. While the pyrimethamine restricted distribution scheme may be unethical and could be bad for public health, this Article addresses the question of whether it violates the antitrust laws. Part II describes the typical distribution systems in the pharmaceutical industry. Part III examines monopoly power and considers whether Daraprim possessed such power. Part IV considers the

1. Kate Gibson, *Martin Shkreli: "I Should've Raised Prices Higher,"* CBS NEWS (Dec. 4, 2015), <http://www.cbsnews.com/news/martin-shkreli-i-shouldve-raised-prices-higher/> [https://perma.cc/TSF9-QAWE].

2. Andrew Pollack & Julie Creswell, *The Mercurial Man Behind the Drug Price Increase that Went Viral*, N.Y. TIMES, Sept. 23, 2015, at B1.

3. Monica V. Mahoney, *New Pyrimethamine Dispensing Program: What Pharmacists Should Know*, PHARM. TIMES (July 17, 2015), <http://www.pharmacytimes.com/contributor/monica-v-golik-mahoney-pharmd-bcps-aq-id/2015/07/new-pyrimethamine-dispensing-program-what-pharmacists-should-know> [https://perma.cc/DHY4-YDGC].

4. See *infra* text accompanying note 116.

second element of monopolization claims, exclusionary conduct, and explores whether Turing engaged in such behavior. Part V then reaches beyond pyrimethamine to offer additional examples of similar conduct. Given that the Federal Trade Commission⁵ and N.Y. Attorney General⁶ are currently conducting antitrust investigations of this behavior, this Article offers a framework for analysis.

II. GENERIC DRUG APPROVAL AND DISTRIBUTION SYSTEMS

Pyrimethamine was originally approved by the FDA in 1953 and was made by its original sponsor, GlaxoSmithKline, and sold for about \$1 per pill.⁷ In 2009, GlaxoSmithKline sold the rights to pyrimethamine to a small, private firm, CorePharma,⁸ which raised the price to \$13.50 per pill.⁹ With about 10,000 prescriptions per year in the United States, sales increased from \$667,000 to \$6.3 million from 2010 to 2011.¹⁰ In 2014, just before Turing bought the rights to pyrimethamine, more than 8,000 prescriptions were written, resulting in sales of \$9.9 million.¹¹

5. *FTC Mounts Antitrust Probe of Shkreli's Ex-Firm Turing: Lawyer*, REUTERS (Jan. 22, 2016), <http://www.nbcnews.com/business/business-news/ftc-mounts-antitrust-probe-shkreli-s-ex-firm-turing-lawyer-n502241> [<https://perma.cc/L4V5-3N2H>].

6. Andrew Pollack, *New York Attorney General Examining Whether Turing Restricted Drug Access*, N.Y. TIMES (Oct. 12, 2015), <http://www.nytimes.com/2015/10/13/business/new-york-attorney-general-examining-if-turing-restricted-drug-access.html?> [<https://perma.cc/C63X-ADYK>].

7. J. Jennings Moss, *With a 5,000 Percent Cost Increase on One Drug, Is This Entrepreneur a Biotech Maverick or Opportunistic Profiteer*, N.Y. BUS. J. (Sept. 21, 2015), <http://www.bizjournals.com/newyork/news/2015/09/21/with-a-5-000-percent-cost-increase-on-one-drug-is.html>.

8. *CorePharma Signs Agreement with GSK to Acquire Dexedrine US NDAs*, PHARMACEUTICAL BUS. REV. (2010), http://contractservices.pharmaceutical-business-review.com/news/corepharma-signs-agreement-with-gsk-to-acquire-dexedrine-us-ndas_251010 [<https://perma.cc/V88L-8X7B>].

9. Jonathan D. Alpern, William M. Stauffer & Aaron S. Kesselheim, *High-Cost Generic Drugs: Implications for Patients and Policymakers*, 371 NEW ENG. J. MED. 1859, 1859–62 (2014).

10. Jennings Moss, *supra* note 7.

11. *Id.* Given that the number of prescriptions did not change during the time period and the costs of production did not increase, it would appear that Turing's profits on this drug also skyrocketed.

Though pyrimethamine was eligible for generic competition by the 1970s, no generic version of the product has yet been approved.¹² The Hatch-Waxman Act of 1984 formalized an abbreviated process for approval of generic drugs based on *in vitro* data as well as pharmacokinetic and pharmacodynamic studies that a manufacturer must conduct between its product and the so-called Reference Listed Drug.¹³ The Reference Listed Drug is the brand-name version designated by the FDA against which a potential generic entrant must test its drug.¹⁴ Upon successful completion of these studies, the FDA can designate a generic drug as bioequivalent and approve its sale in the market, which will then occur as long as the brand manufacturer has no patents or market exclusivities in place.¹⁵ The version of pyrimethamine now owned by Turing is the Reference Listed Drug against which generic manufacturers must test their products to be certified as bioequivalent.¹⁶

Completing bioequivalence studies therefore requires generic manufacturers to obtain samples of the brand-name Reference Listed Drug. Generic manufacturers do that by directly contacting the brand-name manufacturer or working through a wholesaler or other middleman.¹⁷ These transactions are completed without a prescription and with supplies shipped in a bulk form suitable for clinical testing rather than patient use.

After a generic drug is approved and made available for sale, state drug product selection laws permit automatic substitution of FDA-certified

12. FDA, *Drugs@FDA: Daraprim*, <http://www.accessdata.fda.gov/scripts/cder/drugsatfda/index.cfm> (search “Daraprim”) (noting that “[t]here are no Therapeutic Equivalents”) (last visited Apr. 24, 2016).

13. See Aaron S. Kesselheim & Jonathan J. Darrow, *Hatch-Waxman Turns 30: Do We Need a Re-designed Approach for the Modern Era?*, 15 YALE J. HEALTH POL’Y L. & ETHICS 293, 302 (2015) (“The ANDA bioequivalence process permitted approval of generic drugs scientifically proven to work similarly well to their brand-name versions without subjecting those generic drugs to the same clinical trial requirements already completed by the brand-name manufacturer.”).

14. 21 C.F.R. § 314.94(a)(3) (2015); see also FDA, ORANGE BOOK PREFACE (2015), <http://www.fda.gov/Drugs/DevelopmentApprovalProcess/ucm079068.htm> [<https://perma.cc/3MGB-EGYW>] (“By designating a single reference listed drug as the standard to which all generic versions must be shown to be bioequivalent, FDA hopes to avoid possible significant variations among generic drugs and their brand name counterpart.”).

15. Kesselheim & Darrow, *supra* note 13, at 303.

16. FDA, Search Results, *Orange Book: Approved Drug Products with Therapeutic Equivalence Evaluations* (Jan. 7, 2016), http://www.accessdata.fda.gov/scripts/cder/ob/docs/obdetail.cfm?Appl_No=008578&TABLE1=OB_Rx [<https://perma.cc/92SW-LUY6>].

17. See, e.g., Pharmaceutical Buyers, MANTA, <http://www.manta.com/c/mmlh5vy/pharmaceutical-buyers> (last visited Feb. 23, 2016) (“We are an International Company that can provide you with an extended variety of pharmaceutical products for a good price.”).

bioequivalent generic drugs at the pharmacy level. Unless the prescription is marked “dispense-as-written” (which occurs about 5 percent of the time), such substitution can occur even if the prescriber writes the name of the brand-name drug.¹⁸ Because of automatic substitution, patients do not know which company has supplied their generic prescription drug, and generic manufacturers compete largely on the basis of the lowest price they can offer suppliers.¹⁹ The prices at which generic drugs are sold are heavily dependent on the number of manufacturers. In fact, FDA studies show that generic drug prices fall to about 52 percent of the brand price when two generic competitors are in the market, 33 percent when there are five, and 13 percent when there are fifteen.²⁰

Most prescription drugs are available through a standard pharmaceutical distribution chain: from manufacturer to wholesaler, then to retail or mail-order pharmacy, and then to consumer.²¹ The goal is to distribute the drug as widely as possible, because widespread distribution tends to increase manufacturers’ revenues by making drugs available to be prescribed to as many people as possible. The parties contract with one another and hand off control of the drug until it reaches the consumer. Atorvastatin (Lipitor), for example, is manufactured by Pfizer, is distributed by wholesalers such as McKesson, and is available through retail pharmacies such as CVS or Walgreens. In this model, Pfizer relinquishes its control of atorvastatin to McKesson, which then sells the drug to its network of retail pharmacies, with the pharmacies then selling the drug to consumers with valid prescriptions from their physicians. Pfizer is not directly involved at the retail level.

Drugs with limited distribution schemes, by contrast, are not available through standard retail or mail-order pharmacies. Instead, the manufacturer eliminates the wholesaler and distributes the drug only through specialty pharmacies selected by the manufacturer. Manufacturer-sponsored programs like Turing’s Daraprim Direct facilitate the distribution of drugs

18. William H. Shrank et al., *The Consequences of Requesting “Dispense as Written”*, 124 AM. J. MED. 309, 309–10 (2011) (identifying five percent of prescriptions as dispense-as-written in large claims database of prescriptions).

19. Kesselheim & Darrow, *supra* note 13, at 313–14 (“The state DPS laws helped lead to rapid uptake of bioequivalent generic drugs in practice without the time and expense needed to encourage physicians to change their prescribing practices.”).

20. FDA, ANALYSIS OF RETAIL SALES DATA FROM IMS HEALTH, IMS NATIONAL SALES PERSPECTIVE, 1999–2004 (2005).

21. KAISER FAMILY FOUNDATION, FOLLOW THE PILL: UNDERSTANDING THE U.S. COMMERCIAL PHARMACEUTICAL SUPPLY CHAIN (2005), http://avalere.com/research/docs/Follow_the_Pill.pdf [https://perma.cc/H3KJ-8YE8].

from specialty pharmacy to patient. For example, mecasermin (Increlex) is a biologic drug manufactured by Ipsen Pharmaceuticals to treat growth failure and severe primary insulin-like growth deficiency.²² Patients must enroll in Ipsen's "Ipsen Cares" program before receiving the drug. Ipsen then coordinates the delivery of the drug through its specialty pharmacy network.²³ Actelion Pharmaceuticals has a similar program called "Actelion Pathways" for its drug iloprost (Ventavis), a treatment for pulmonary arterial hypertension.²⁴ In this case, physicians must enroll patients in this program through the manufacturer for a specialty pharmacy to deliver the drug.²⁵

When safety issues arise in the clinical trials supporting approval of a drug, the FDA may require the use of Risk Evaluation and Mitigation Strategies (REMS) to ensure that a drug's benefits outweigh its risks.²⁶ The FDA can require REMS that take the form of medication guides, patient package inserts, communication plans, or elements to assure safe use (ETASU) (with this last category including restrictions on how drugs are distributed to patients).²⁷ Restricted distribution in these cases may be justified because it allows manufacturers to track prescriptions and monitor patients. For example, lenalidomide (Revlimid), a treatment for multiple myeloma, is believed to cause serious birth defects. To avoid embryo-fetal exposure, it is available only through restricted distribution to ensure that the drug is prescribed only to women who are not pregnant or trying to conceive.²⁸

22. FDA, HIGHLIGHTS OF PRESCRIBING INFORMATION: MECASERMIN RECOMBINANT (INCRELEX) (2014), http://www.accessdata.fda.gov/drugsatfda_docs/label/2014/021839s0161bl.pdf [<https://perma.cc/5RRM-VZTA>].

23. IPSEN CARES (2016), <http://www.ipsencares.com/#about-ipsen-cares> [<https://perma.cc/XZ5K-FMEG>].

24. FDA, HIGHLIGHTS OF PRESCRIBING INFORMATION: ACTELION PHARMS LTD., ILOPROST (VENTAVIS) (2013), http://www.accessdata.fda.gov/drugsatfda_docs/label/2013/021779s0141bl.pdf [<https://perma.cc/XT8U-6PW6>].

25. *Working with a Specialty Pharmacy*, ACTELION PATHWAYS, <https://www.actelionpathways.com/ventavis-and-specialty-pharmacies> [<https://perma.cc/V32J-HS72>] (last visited Feb. 24, 2016).

26. *See* 21 U.S.C. § 355-1(a)(1) (2012).

27. FDA, CENTER FOR DRUG EVALUATION AND RESEARCH, HOW TO OBTAIN A LETTER FROM FDA STATING THAT BIOEQUIVALENCE STUDY PROTOCOLS CONTAIN SAFETY PROTECTIONS COMPARABLE TO APPLICABLE REMS FOR RLD: GUIDANCE FOR INDUSTRY (Dec. 2014), <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM425662.pdf> [<https://perma.cc/KP47-NCBV>].

28. Amee Sarpatwari et al., *Using a Drug-Safety Tool to Prevent Competition*, 370 NEW ENG. J. MED. 1476, 1477 (2014) (listing key dangerous side effects of lenalidomide).

At the same time, however, limiting sales of a product through one particular wholesaler also gives the manufacturer complete control over the distribution chain. In public forums, some manufacturers of limited distribution drugs have emphasized that they can provide patient-centered programs as part of their restricted distribution schemes. These programs use the narrow distribution pool to monitor patients and provide certain types of adherence support, such as assistance with refilling, the ability to ask questions to manufacturer representatives, and connecting patients with one another to provide social support.²⁹ But limited distribution also can allow brands to restrict access to samples needed by generics in their bioequivalence studies. In particular, moving supply through a single source can allow the brand to take steps to prevent the supply of the product to a generic that might otherwise have gone to pharmacies for use in filling patient prescriptions. As a result, limited distribution systems create a market environment in which anticompetitive behavior can thrive.

III. MONOPOLY POWER

The relevant antitrust law in considering the actions of Turing is monopolization, which focuses on the conduct of a single company. To prove a monopolization claim, a plaintiff must show monopoly power and exclusionary conduct. This Part will analyze monopoly power and Part IV will address exclusionary conduct.

Monopoly power has been defined as “the power to control prices or exclude competition.”³⁰ It can be shown in one of two ways. First, it can be proved indirectly by examining a defendant’s market share along with barriers to entry that could entrench that market position.³¹ Second, it can be proved directly,³² such as when a brand firm is able to “maintain the price of [a drug] at supracompetitive levels without losing substantial sales.”³³ In addition to these antitrust requirements, Part III addresses the most potent

29. Yifei Liu et al., *Greater Refill Adherence to Adalimumab Therapy for Patients Using Specialty Versus Retail Pharmacies*, 27 *ADVANCES IN THERAPY* 523, 523–30 (2010).

30. *United States v. E.I. duPont de Nemours & Co.*, 351 U.S. 377, 391 (1956).

31. HERBERT HOVENKAMP, *FEDERAL ANTITRUST POLICY: THE LAW OF COMPETITION AND ITS PRACTICE* ¶ 6.2b, at 359–60 (5th ed. 2016).

32. ABA SECTION OF ANTITRUST LAW, *ANTITRUST LAW DEVELOPMENTS* 69–70 (7th ed. 2012) (noting that “direct proof has provided the basis for findings of substantial anticompetitive effects in some prominent cases”).

33. *In re Nexium (Esomeprazole) Antitrust Litig.*, 968 F. Supp. 2d 367, 387 (D. Mass. 2013); *see also In re Aggrenox Antitrust Litig.*, 94 F. Supp. 3d 224, 241 (D. Conn. 2015).

argument against monopoly power in this case: the existence of an inexpensive “compounded version” of the drug.

A. INDIRECT PROOF

Monopoly power can be demonstrated indirectly by defining a relevant market and examining the company’s share of the market. Courts regularly hold that a 90 percent market share supports market power, with several courts finding a 75 percent share to be sufficient.³⁴

Evidence that Turing has 100 percent of the relevant market is provided by the lack of effective, FDA-approved substitutes. Pyrimethamine is part of all widely accepted first-line therapeutic regimens for toxoplasmosis.³⁵ While toxoplasmosis has been treated without pyrimethamine and with alternative treatments, such as sulfamethoxazole-trimethaprim (Bactrim) and clindamycin (Cleocin), the efficacy of these approaches is currently based only on case reports³⁶ and other less rigorous data.³⁷ In fact, the American Society of Microbiology warned that the 5000 percent price increase would “negatively impact both health care costs and individual patient treatments.”³⁸ Nor, as discussed in detail below,³⁹ is a compounded version an effective substitute for pyrimethamine.

Regulatory barriers to entry cement the effect of this high market share. As discussed in greater detail below, generics can enter the U.S. market only after receiving FDA approval. Turing’s restriction of its distribution system entrenches its monopoly power by preventing generics from obtaining the samples needed for bioequivalence testing.

34. HOVENKAMP, *supra* note 31, ¶ 6.2a, at 357.

35. Sara Fazio, *Toxoplasmosis*, NEW ENG. J. MED. BLOG (Feb. 23, 2012), <http://blogs.nejm.org/now/index.php/toxoplasmosis/2012/02/23/> [<https://perma.cc/P66V-DWJS>].

36. *See, e.g.*, Deepak Madi et al., *Successful Treatment of Cerebral Toxoplasmosis with Clindamycin: A Case Report*, 27 OMAN MED. J. 411 (2012).

37. Pollack, *supra* note 6.

38. Memorandum from the Democratic Staff to Democratic Members of the Full House Comm. on Oversight and Gov’t Relations, at 5 (Feb. 2, 2016), <https://democrats-oversight.house.gov/sites/democrats.oversight.house.gov/files/documents/Memo%20on%20Turing%20Documents.pdf> [<https://perma.cc/CA9E-MFCF>].

39. *See infra* Section III.C.

B. DIRECT PROOF

Direct proof of monopoly power consists of observable effects on the market such as a price increase or output reduction.⁴⁰ Turing's conduct has revealed both types of direct evidence.

To begin, Turing's price increase has received unparalleled attention. Even though there has not been an increase in the costs of producing pyrimethamine (which costs pennies per pill to manufacture⁴¹), Turing has increased the price 5000 percent. In addition, Turing has been able to maintain that increase despite public outrage and substantial attention from the lay press, Congress, and Presidential candidates.⁴² Shkreli initially announced that Turing would lower the price for the drug in response to the negative publicity but later revealed that this reduction would be only 10 percent.⁴³ Ultimately, Shkreli decided not to lower the price at all, instead offering free samples, rebates to hospitals, and smaller bottle sizes.⁴⁴ Given the barriers to entry imposed by obtaining FDA review, the high prices will likely be maintained over an extended period of time.⁴⁵

Documents provided to the House Committee on Oversight and Government Reform offer numerous examples of price increases including patient copays in the thousands of dollars. The Director of Specialty Pharmacy Development at Walgreens recounted anecdotes of patients having difficulty obtaining pyrimethamine, including one who was forced to make a \$6,000 copay.⁴⁶ An internal presentation reported that “[p]atients with commercial/private insurance [are] experiencing increased co-pays,

40. *Broadcom Corp. v. Qualcomm Inc.*, 501 F.3d 297, 307 (3d Cir. 2007).

41. See Karthick Arvinth, *Daraprim: Generic Version of Drug Costs Less than £0.07 in India*, INTERNATIONAL BUSINESS TIMES (Sept. 25, 2015), <http://www.ibtimes.co.uk/daraprim-like-drug-costs-less-0-07-india-1521144> [<https://perma.cc/4RAZ-TRJH>].

42. Andrew Pollack, *Drug Goes From \$13.50 a Tablet to \$750, Overnight*, N.Y. TIMES, Sept. 21, 2015, at B1.

43. Andrew Pollack, *Turing Commits to Modest Price Reduction on a Drug*, N.Y. TIMES, Nov. 4, 2015, at B3.

44. Sam Thielman, *Martin Shkreli Walks Back on Pledge to Lower Price of HIV Drug Daraprim*, THE GUARDIAN (Nov. 25, 2015), <http://www.theguardian.com/business/2015/nov/25/martin-shkreli-hiv-drug-daraprim-turing> [<https://perma.cc/XLX5-WT7X>]. Turing documents reveal a methodical campaign to increase price by anticipating the reactions of HIV/AIDS groups and doctors. See Comm. Memorandum, *supra* note 38, at 3 (“Physician community less sensitive to price increases, but need to determine the price point at which payers start to increase cost-sharing with patients, which could result in physician switching.”).

45. See, e.g., *Star Fuel Marts, LLC v. Sam’s E., Inc.*, 362 F.3d 639, 654 (10th Cir. 2004).

46. Comm. Memorandum, *supra* note 38, at 4.

delays in claims approval[,] and rejections,” with one facing a copy of \$16,830.⁴⁷

Output reductions are another direct indicator of monopoly power. After pyrimethamine’s price increase, hospitals complained that they were not able to obtain the drug.⁴⁸ Turing’s own press release conceded that hospitals and clinics “were having trouble accessing the product.”⁴⁹

The combination of a price increase and output reduction is a hallmark of monopoly power, and the Democratic Staff memorandum synthesizing 250,000 pages of Turing documents revealed just such effects:

Daraprim has now become prohibitively expensive, hospital budgets are straining under the huge cost increases, patients are being forced to pay thousands of dollars in co-pays and are experiencing major challenges obtaining access to the drug, and physicians are considering using alternative therapies.⁵⁰

C. MONOPOLY POWER NOT NEGATED BY COMPOUNDED VERSION

On October 22, 2015, Imprimis Pharmaceuticals announced that it was planning to make available for \$1 a compounded coformulation of pyrimethamine and leucovorin (a folic acid derivative usually coprescribed with pyrimethamine as a separate pill to help protect against its side effects of bone marrow suppression).⁵¹ Thus, a counterargument to the conclusion of monopoly power would be that the compounded version serves as a substitute. Such an argument would point to certain patients taking this version instead of the FDA-approved version sold by Turing. If patients are in fact able to substitute the compounded version, then that could conceivably show a lack of market power.

47. *Id.* at 5.

48. Letter from Stephen B. Calderwood & Adaora Adimora to Tom Evegán & Kevin Bernier (Sept. 8, 2015), <http://www.hivma.org/uploadedFiles/HIVMA/HomePageContent/PyrimethamineLetterFINAL.pdf> [<https://perma.cc/8TZ9-KY3V>].

49. *Press Release: Important News about Daraprim (pyrimethamine)*, TURING PHARMACEUTICALS (Sept. 18, 2015), <http://www.turingpharma.com/media/press-release?headline=important-news-about-daraprim%25c2%25ae-%28pyrimethamine%29> [<https://perma.cc/8V3P-V4QS>].

50. Comm. Memorandum, *supra* note 38, at 1.

51. *Imprimis Pharmaceuticals to Make Compounded and Customizable Formulation of Pyrimethamine and Leucovorin Available for Physicians to Prescribe for Their Patients as an Alternative to Daraprim*, PR NEWSWIRE (Oct. 22, 2015), <http://www.prnewswire.com/news-releases/imprimis-pharmaceuticals-to-make-compounded-and-customizable-formulation-of-pyrimethamine-and-leucovorin-available-for-physicians-to-prescribe-for-their-patients-as-an-alternative-to-daraprim-300164514.html> [<https://perma.cc/5MFQ-VCMH>].

Such an argument is not persuasive. Imprimis's combination pill does not address the problem of costly pyrimethamine because the compounded drug is not an effective market substitute. Compounded drugs are synthesized at specially licensed pharmacies to respond to individual requests for variations of particular active ingredients that cannot be obtained through FDA-approved channels.⁵² Compounded drugs can include new formulations of products. For example, a compounding pharmacy might create a lozenge version of a medication available in pill form for a patient who has problems swallowing pills or a different concentration of an intravenous drug sold in only one strength.⁵³

Compounding pharmacies have historically not been permitted to distribute their products in bulk.⁵⁴ But a recent provision of the Food, Drug, and Cosmetics Act (FDCA) allows compounding pharmacies to register as outsourcing facilities, which permits the sale of compounded drugs in bulk and requires manufacturers to comply with current Good Manufacturing Practices.⁵⁵ Imprimis plans to register at least one of its compounding pharmacies as an outsourcing facility, which would allow mass production of pyrimethamine/leucovorin and sales to hospitals and physicians.⁵⁶

While compounded drugs produced by outsourcing facilities may resemble FDA-approved drugs, they are not. For starters, compounded drugs by definition cannot be a direct substitute for FDA-approved drugs.

52. Kevin Outterson, *Regulating Compounding Pharmacies after NECC*, 367 NEW ENG. J. MED. 1969, 1971 (2012) (“[T]raditional compounding was limited to a pharmacist or a physician serving a specific patient.”).

53. See, e.g., Loyd V. Allen Jr., *Troches and Lozenges*, 4 SECUNDUM ARTEM 2, <http://www.perrigo.com/business/pdfs/Sec%20Artem%204.2.pdf> [<https://perma.cc/SQT4-68Z8>] (“Lozenges, or troches, are experiencing renewed popularity as a means of delivering many different drug products. They are used for patients who cannot swallow solid oral dosage forms . . .”).

54. Outterson, *supra* note 52, at 1970 (describing 2002 FDA compliance policy guide not permitting use of “commercial-scale manufacturing or testing equipment for compounding drug products”).

55. See Food Drug and Cosmetic Act §§ 503A, 503B. Section 503B, which created a new category of compounders called “outsourcing facilities,” was added to the Food Drug and Cosmetic Act by the Drug Quality and Security Act in 2013. See Pub. L. No. 113-54, § 102, 127 Stat. 587, 588 (2013) (amending 21 U.S.C. § 351 et seq.).

56. *Imprimis Pharmaceuticals Announces Plans to Register Its Texas Pharmacy with the FDA as an Outsourcing Facility*, PR NEWSWIRE (Oct. 29, 2015), <http://www.prnewswire.com/news-releases/imprimis-pharmaceuticals-announces-plans-to-register-its-texas-pharmacy-with-the-fda-as-an-outsourcing-facility-300168448.html> [<https://perma.cc/ZGU9-CNS7>]; FDA, *Outsourcing Facilities* (Oct. 6, 2015), <http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/PharmacyCompounding/ucm393571.htm> [<https://perma.cc/9DXP-8GPD>].

Indeed, one restriction on compounded drugs is that they not be “essentially a copy of a commercially available drug product,”⁵⁷ which is why Imprimis’s version of pyrimethamine contains leucovorin.

More important, the FDA does not verify the safety, effectiveness, or manufacturing quality of compounded drugs in traditional compounding pharmacies. Instead, they are regulated by state pharmacy boards.⁵⁸ Though compounding pharmacies registered as outsourcing facilities are inspected by the FDA and must report adverse events,⁵⁹ FDA regulation of compounding pharmacies has traditionally been secondary to oversight by state inspectors,⁶⁰ and there can be substantial state-to-state variations in state authority and resources dedicated to this area.⁶¹

In response to a 2012 meningitis outbreak originating from the New England Compounding Center, the FDA has increased oversight of compounding pharmacies. Still, the agency cautions that poor quality-control practices may result in compounded drugs that are “sub- or super-potent, contaminated, or otherwise adulterated.”⁶² Patients subject themselves to risk when they “use ineffective compounded drugs instead of FDA-approved drugs that have been shown to be safe and effective.”⁶³ With

57. 21 U.S.C. § 353a.

58. Roy Guharoy et al., *Compounding Pharmacy Conundrum: “We Cannot Live without Them But We Cannot Live with Them” According to the Present Paradigm*, 143 CHEST 896, 897 (2013).

59. FDA, GUIDANCE FOR INDUSTRY, ADVERSE EVENT REPORTING FOR OUTSOURCING FACILITIES UNDER SECTION 503B OF THE FEDERAL FOOD, DRUG, AND COSMETIC ACT (Oct. 2015), <http://www.fda.gov/ucm/groups/fdagov-public/@fdagov-drugs-gen/documents/document/ucm434188.pdf> [<https://perma.cc/DR9D-MQL8>].

60. Steven K. Galston, *Federal and State Role in Pharmacy Compounding and Reconstitution: Exploring the Right Mix to Protect Patients*, TESTIMONY BEFORE HEALTH, EDUCATION, LABOR, AND PENSIONS SENATE COMM. (Oct. 23, 2003), <http://www.fda.gov/NewsEvents/Testimony/ucm115010.htm> [<https://perma.cc/U8MY-4TBF>] (“FDA has historically exercised its enforcement discretion in a manner that defers to the states, as the regulators of the practice of pharmacy, to serve as the primary regulators of the practice of pharmacy compounding.”).

61. Jennifer Gudeman et al., *Potential Risks of Pharmacy Compounding*, 13 DRUGS R&D 1 (2013) (“The FDA regulates and regularly inspects pharmaceutical manufacturing facilities to ensure compliance with GMPs. In contrast, pharmacies are primarily under the authority of state Boards of Pharmacy . . . and only undergo FDA inspections in rare instances. As a result, there is less assurance of consistent quality for compounded preparations than there is for FDA-approved drugs.”).

62. FDA, *Compounding and the FDA: Questions and Answers*, <http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/PharmacyCompounding/ucm339764.htm> [<https://perma.cc/AP9N-2QGJ>] (last visited June 7, 2015).

63. *Id.*

regards to efficacy, there are no requirements for clinical testing of the potency of nonsterile compounded drugs (e.g., tablets, creams, lozenges).⁶⁴

Even though third-party testing of Imprimis's pyrimethamine/leucovorin revealed that the drug met FDA-recognized potency standards,⁶⁵ safety remains a concern, as contaminated compounded products have been implicated in public health crises.⁶⁶ A recent review evaluated 11 infectious outbreaks caused by contaminated compounded medications, affecting 207 patients and causing 17 deaths, and identified inadequate regulatory controls as the major underlying cause.⁶⁷ Not included in this total was a 2012 epidemic caused by fungal contamination of an injectable steroid prepared by a compounding pharmacy, which resulted in 749 serious infections in 20 states, including 229 cases of meningitis and 61 deaths.⁶⁸ And one study concluded that 41 percent of doctors considered the lack of FDA approval of a drug preventing preterm delivery as a deterrent to prescribing the medication, with 39 percent having professional liability concerns prescribing the compounded drug.⁶⁹

While efficacy and safety risks vary by the particular compounder and the specific product, these considerations make compounded drugs unlikely to achieve the same level of widespread acceptance among physicians and patients as FDA-approved drugs.

64. See *Pharmaceutical Compounding—Nonsterile Preparations*, U.S. PHARMACOPEIAL CONVENTION (Jan. 1, 2014), http://www.usp.org/sites/default/files/usp_pdf/EN/gc795.pdf [<https://perma.cc/MAV5-FFND>] (describing lack of potency testing in rules for creating nonsterile compounds); cf. *Pharmaceutical Compounding—Sterile Preparations*, U.S. PHARMACOPEIAL CONVENTION, at 42 (June 1, 2008), <http://www.pbm.va.gov/linksotherresources/docs/USP797PharmaceuticalCompoundingSterileCompounding.pdf> [<https://perma.cc/T4MY-GSJU>].

65. *Imprimis Pharmaceuticals Reports Results of Independent Third Party Potency Analysis of Its Compounded Pyrimethamine and Leucovorin Capsules*, PR NEWswire (Dec. 11, 2015), <http://www.prnewswire.com/news-releases/imprimis-pharmaceuticals-reports-results-of-independent-third-party-potency-analysis-of-its-compounded-pyrimethamine-and-leucovorin-capsules-300191649.html> [<https://perma.cc/2CN6-AXGM>].

66. Outterson, *supra* note 52, at 1971 (describing New England Compounding crisis).

67. C. Catherine Staes et al., *Description of Outbreaks of HealthCare Associated Infections Related to Compounding Pharmacies, 2000–2012*, 70 AM. J. HEALTH SYS. PHARM. 1301 (2013).

68. Rachel M. Smith et al., *Fungal Infections Associated with Contaminated Methylprednisolone Injections*, 369 NEW ENG. J. MED. 1598 (2013).

69. Andrei Rebarber et al., *A National Survey Examining Obstetrician Perspectives on Use of 17-Alpha Hydroxyprogesterone Caproate Post-US FDA Approval*, 33 CLINICAL DRUG INVESTIGATION 571, 573 (2013), <http://www.ncbi.nlm.nih.gov/pubmed/23800978> [<https://perma.cc/K56E-6BAK>].

In addition, it is often difficult for patients to have compounded drugs covered by their insurance company.⁷⁰ When pharmaceutical benefits manager ExpressScripts announced that it would reimburse Imprimis's version of pyrimethamine/leucovorin, the announcement was remarkable enough that it made the national news.⁷¹ ExpressScripts admitted that it would also continue to cover a "general prescription" of the drug and that patients could only obtain Imprimis's version by having their physician send a special prescription directly to Imprimis.⁷² Shkreli himself asserted that Imprimis's compounded drug "isn't really an alternative" to Turing's pyrimethamine.⁷³ In short, while compounded drugs may be sufficient for certain individual patients, they are not substitutes in the market as a whole.⁷⁴ Given that there are no FDA-approved substitutes for Turing's pyrimethamine and that Turing has been able to increase price significantly and reduce output, it has monopoly power.

IV. EXCLUSIONARY CONDUCT

To bring a successful monopolization claim, a plaintiff must show not only monopoly power but also exclusionary conduct. This Part first offers an overview of the case law on exclusionary conduct before more specifically exploring the law relating to the denial of drug samples in the context of FDA-required safety programs. It then applies this case law to Turing, showing how the restriction of its distribution system reveals exclusionary conduct.

70. See Ed Silverman, *Express Scripts Ends Coverage for 1,000 Compound Drug Ingredients*, WALL ST. J.: PHARMALOT (July 1, 2014), <http://blogs.wsj.com/pharmalot/2014/07/01/express-scripts-ends-coverage-for-1000-compound-drug-ingredients/> [https://perma.cc/B9VQ-ACXU].

71. Jonathan D. Rockoff, *Express Scripts Turns to a Compounder to Avoid a Turing Drug*, WALL ST. J. (Dec. 1, 2015), <http://www.wsj.com/articles/express-scripts-seeks-lower-price-alternative-to-daraprim-1448946061> [https://perma.cc/Y4JF-VHMF].

72. *Id.*

73. Lucas Matney, *Turing CEO Defends \$750 Pill in Reddit AMA, Says Scandal Has Been "Best Possible Way to Get Girls,"* TECHCRUNCH (Oct. 25, 2015), <http://techcrunch.com/gallery/turing-ceo-defends-750-pill-in-reddit-ama-says-scandal-has-been-best-possible-way-to-get-girls/slide/20/> [https://perma.cc/R6CL-EMTA].

74. Jeremy A. Greene et al., *Role of the FDA in Affordability of Off-Patent Pharmaceuticals*, 315 JAMA 461–62 (Jan. 4, 2016), <http://jama.jamanetwork.com/article.aspx?articleid=2480263&resultClick=3> [https://perma.cc/BCD4-7DJ6].

A. MONOPOLIZATION CASE LAW

To be liable for illegal monopolization, a company not only must have monopoly power but also must engage in exclusionary conduct. Courts often distinguish between the “willful acquisition or maintenance of [monopoly] power” and “growth or development as a consequence of a superior product, business acumen, or historic accident.”⁷⁵

This test is more difficult to apply than to state. Certain cases have served as landmarks to guide analysis. For example, in *Aspen Skiing Co. v. Aspen Highlands Skiing Corp.*, the owner of three downhill skiing facilities in Aspen, Colorado failed to offer a justification for withdrawing from a joint ticketing arrangement with the owner of the only other facility in the area. The Supreme Court defined exclusionary conduct as that which “tends to impair the opportunities of rivals” and which “either does not further competition on the merits or does so in an unnecessarily restrictive way.”⁷⁶ The Court found that the monopolist was guilty of anticompetitive conduct because it was willing to forego ticket sales and sacrifice profits to harm its smaller competitor.⁷⁷ As applied by commentators, this profit-sacrifice test offers a defendant-friendly approach that only punishes activity that has no justifiable reason other than harming competitors.⁷⁸

In a second classic case, *Otter Tail Power Co. v. United States*, the Supreme Court required a company to share electric power transmission with rivals.⁷⁹ The company “was already in the business of providing a service to certain customers,” and thus could not “refuse[] to provide the same service to certain other customers.”⁸⁰ In particular, there were “no engineering factors that prevented Otter Tail from selling power at

75. *United States v. Grinnell Corp.*, 384 U.S. 563, 570–71 (1966).

76. 472 U.S. 585, 605 n.32 (1985).

77. *Id.* at 608.

78. E.g., A. Douglas Melamed, *Exclusive Dealing Agreements and Other Exclusionary Conduct—Are There Unifying Principles?*, 73 ANTITRUST L.J. 375, 392–93 (2006) (“anticompetitive intent” of firm willing to sacrifice profits can be “unambiguously inferred”); Gregory J. Werden, *Identifying Exclusionary Conduct under Section 2: The “No Economic Sense” Test*, 73 ANTITRUST L.J. 413, 415 (2006) (the test’s application “could not be simpler if . . . the conduct cannot possibly confer an economic benefit on the defendant other than by eliminating competition”); Steve D. Shadowen et al., *Anticompetitive Product Changes in the Pharmaceutical Industry*, 41 RUTGERS L.J. 1, 75–76 (2009) (profit sacrifice leads to natural inference that actor “was aware of and motivated solely to achieve that reduction”).

79. *Otter Tail Power Co. v. United States*, 410 U.S. 366 (1973).

80. *Verizon Commc’ns Inc. v. Law Offices of Curtis V. Trinko, LLP*, 540 U.S. 398, 410 (2004).

wholesale to those towns that wanted municipal plants or [transferring] the power.”⁸¹ Rather, its “refusals to sell at wholesale or to [transfer] were solely to prevent municipal power systems from eroding its monopolistic position.”⁸²

A third case underscored the importance of an effective regulatory regime that covered the conduct, reducing the need for antitrust. In *Verizon Communications v. Trinko*, the Supreme Court held that the Telecommunications Act of 1996 promoted competition by breaking up local phone service monopolies and effectively did so by imposing a regulatory regime that included penalties and reporting requirements.⁸³ The Supreme Court distinguished the *Aspen Skiing* and *Otter Tail* cases by noting that the defendants in those cases offered ski lift tickets and power transmission, respectively, which were services already available to the public.⁸⁴ By contrast, Verizon was required to share unbundled network elements, a “brand new” type of service that “exist[ed] only deep within the bowels” of the company.⁸⁵ These network elements were “offered not to consumers but to rivals, and at considerable expense and effort,” which played a role in the dismissal of *Trinko*’s claim.⁸⁶ The Court also worried about requiring a firm to share with its rivals, as such a remedy would “require[] antitrust courts to act as central planners” and could “facilitate the supreme evil of antitrust: collusion.”⁸⁷

Courts since *Trinko* have been skeptical of refusal-to-deal cases, worrying about the effects of forcing a company to collaborate with rivals. But as we discuss below, Turing’s conduct is closer to that in *Aspen Skiing* and *Otter Tail* than *Trinko*. The next section turns closer to the facts at issue with pyrimethamine.

B. RISK EVALUATION AND MITIGATION STRATEGY (REMS) CASE LAW

As discussed above, when safety issues arise in the clinical trials supporting approval of a drug, the FDA may require the use of Risk Evaluation and Mitigation Strategies (REMS) to ensure that a drug’s benefits outweigh its risks.⁸⁸ Although no antitrust case has analyzed issues

81. *Otter Tail Power*, 410 U.S. at 378.

82. *Id.*

83. 540 U.S. at 410–11.

84. *Id.*

85. *Id.*

86. *Id.*

87. *Id.* at 408.

88. See *supra* text accompanying notes 26–28.

of the restricted distribution of pharmaceuticals outside the REMS setting, several cases have considered similar issues in the REMS context. Decisions in these cases have revealed that refusing to sell pharmaceutical samples can constitute exclusionary conduct.

In the first case, Lannett sued Celgene, seeking samples of thalidomide (Thalomid), the infamous drug that was found in the 1960s to cause devastating birth defects when used as an anti-nauseant but was later found to be an effective treatment for leprosy and multiple myeloma.⁸⁹ Celgene had been selling the drug under an FDA-approved special distribution scheme called System for Thalidomide Education and Prescribing Safety (STEPS) that was designed to prevent the drug from being inadvertently prescribed to pregnant women.⁹⁰ STEPS included “prescriber and pharmacy certification, patient registration, and limitations on drug dispensing” that required patients and prescribers to complete a phone survey identifying risk-increasing behavior before a prescription could be issued.⁹¹ In denying the defendants’ motion to dismiss, the District of New Jersey court ruled that prior cases that “have considered the scope of the affirmative duty to deal suggest that a ‘prior course of dealing’ is relevant but not dispositive in determining whether such a duty applies.”⁹² In addition, the court made clear that “the question of whether a defendant sold its product at retail . . . is relevant to determining whether Section . . . 2 liability applies.”⁹³

In a second case, Actelion filed a declaratory judgment action against Apotex, Roxane, and Actavis to affirm that it did not have an obligation to supply samples of bosentan (Tracleer), a vasodilation drug used to treat pulmonary arterial hypertension.⁹⁴ Actelion argued that “its distribution of bosentan [was] restricted to pharmacies certified under the Tracleer Access Program, which require[d] education, counseling, and monthly follow-up of enrolled patients for liver function and pregnancy tests,” and thus that it “could not provide potential competitors with samples of the drug.”⁹⁵ In

89. Erin Coe, *Lannett Cuts Deal with Celgene in Thalomid Antitrust Case*, LAW360 (Dec. 7, 2011), <http://www.law360.com/articles/291483/lannett-cuts-deal-with-celgene-in-thalomid-antitrust-case> [<https://perma.cc/DSZ2-HB55>]; Sarpatwari et al., *supra* note 28, at 1477 (describing new uses of thalidomide).

90. Coe, *supra* note 89.

91. *Id.*

92. Transcript of Oral Opinion at 12–13, *Mylan Pharms. Inc. v. Celgene Corp.*, No. 2:14-cv-02094-ES-MAH (D.N.J. Dec. 22, 2014).

93. *Id.* at 12.

94. Sarpatwari et al., *supra* note 28, at 1476–77 (describing controversy over generic manufacturers’ ability to access samples of bosentan).

95. Sarpatwari et al., *supra* note 28.

announcing that it would allow the case against Actelion to proceed, the district court noted that the Supreme Court's refusal-to-deal decisions were "fact-specific" and "industry-specific" and that the generics "alleged a profit motive which did not exist in *Trinko*."⁹⁶ In addition, the court observed that "the FDA does not have the regulatory power to compel samples and . . . there is no other potential remedy to a defendant suffering anticompetitive conduct in that regulatory scheme."⁹⁷

In a third case, Mylan sued Celgene, challenging its denial of a follow-on variation of thalidomide, lenalidomide (Revlimid), which was sold under a similar program to STEPS.⁹⁸ Even after the FDA determined that Mylan's testing safety protocols were acceptable, Celgene stalled Mylan's efforts to obtain samples by imposing unnecessary requests for additional information. The court found that the plaintiffs had successfully pled a monopolization case by pointing to Celgene's lack of a "legitimate business reason" for its actions, which allegedly were "solely motivated by its goal to obtain long-term anticompetitive gain."⁹⁹

As of this writing, the *Mylan v. Celgene* case is ongoing, with the other two cases having settled after the courts refused to dismiss the plaintiffs' claims.¹⁰⁰ As a result, no final decisions on these issues have been rendered. But the cases chart a potential path to liability for a brand manufacturer's refusal to provide samples to generic rivals.

96. Transcript of Motions Hearing at 115, *Actelion Pharm. Ltd. v. Apotex, Inc.*, No. 1:12-cv-05743 (D.N.J. Oct. 17, 2013).

97. *See id.* at 115–16. In contrast to the lack of FDA authority, the Court in *Trinko* highlighted the Federal Communications Commission's ability to control incumbent telephone carriers' entry into the long-distance market and its enforcement through oversight, penalties, and the revocation of approval to enter the long-distance market. *Verizon Commc'ns Inc. v. Law Offices of Curtis V. Trinko, LLP*, 540 U.S. 398, 412 (2004).

98. *See* Transcript of Oral Opinion, *Mylan Pharms. v. Celgene Corp.*, No. 2:14-cv-02094-ES-MAH (D.N.J., Dec. 22, 2014).

99. *Id.* at 17. In a different setting, in which plaintiffs alleged that they were not able to obtain the samples needed for bioequivalency testing but that the brand firm refused to cooperate in setting up an FDA-required Single Shared REMS program (SSRS), the court dismissed the case. *See In re Suboxone Antitrust Litig.*, 64 F. Supp. 3d 665 (E.D. Pa. 2014).

100. Kurt R. Karst, *Another REMS Antitrust Lawsuit: Mylan Sues Celgene over THALOMID and REVLIMID to Obtain Drug Product Sample*, FDA LAW BLOG (Apr. 4, 2014), http://www.fdalawblog.net/fda_law_blog_hyman_phelps/2014/04/another-rems-anti-trust-lawsuit-mylan-sues-celgene-over-thalomid-and-revlimid-to-obtain-drug-product-.html [<https://perma.cc/6QHY-8PV7>].

C. APPLICATION TO TURING

In considering whether Turing's refusal to provide samples constitutes exclusionary conduct, the regulatory background is essential. The Supreme Court in *Trinko* explained that "antitrust analysis must always be attuned to the particular structure and circumstances of the industry at issue."¹⁰¹ In particular, courts must take "careful account" of "the pervasive federal and state regulation characteristic of the industry," and the analysis must "recognize and reflect the distinctive economic and legal setting of the regulated industry to which it applies."¹⁰²

A central objective of the Hatch-Waxman Act is to encourage generic entry.¹⁰³ Congress sought to achieve this goal through several mechanisms, including formalizing the expedited pathway and allowing generic firms to experiment on a brand firm's drug before the end of the patent term (an otherwise impermissible use).¹⁰⁴ As previously discussed, the Hatch-Waxman scheme allows generic manufacturers to earn abbreviated approvals if they can show that their drugs are bioequivalent to the Reference Listed Drug by testing samples acquired from distributors or wholesalers.¹⁰⁵

This crucial element of competition in the pharmaceutical marketplace is possible only if the generic has access to the brand firm's samples.¹⁰⁶ But as evidenced above, monopolists can improperly design their restricted distribution systems to prevent distributors and wholesalers from selling the drug to competing manufacturers. And the brand itself then can refuse to sell to the generic. The combination of the restricted distribution system and the brand's refusal to deal with the generic would result in the generic lacking access to the samples needed for testing and not being able to

101. *Trinko*, 540 U.S. at 411.

102. *Id.*

103. See, e.g., Michael A. Carrier, *Unsettling Drug Patent Settlements: A Framework for Presumptive Illegality*, 108 MICH. L. REV. 37, 41–43 (2009).

104. See 35 U.S.C. § 271(e)(1) (2012); *Eli Lilly & Co., v. Medtronic, Inc.*, 496 U.S. 661, 669–70 (1990) (allowing experimentation before end of patent term would prevent "unintended distortion" of patent laws that would extend "de facto monopoly"); FTC, *GENERIC DRUG ENTRY PRIOR TO PATENT EXPIRATION: AN FTC STUDY 5* (2002), https://www.ftc.gov/sites/default/files/documents/reports/generic-drug-entry-prior-patent-expiration-ftc-study/genericdrugstudy_0.pdf [<https://perma.cc/4935-QRHF>].

105. See Lauren Battaglia, *Risky Conduct with Risk Mitigation Strategies? The Potential Antitrust Issues Associated with REMS*, ANTITRUST HEALTH CARE CHRONICLE 26, 28 (Mar. 2013).

106. Kesselheim & Darrow, *supra* note 13, at 340–41.

demonstrate the bioequivalence needed to file an application. This could lead to a significant weakening of the regulatory regime.¹⁰⁷

Restricting the typical expansive distribution scheme also tends to involve a sacrifice of the brand's profits. As mentioned above, most prescription drugs are available through a standard pharmaceutical distribution chain: from manufacturer to wholesaler, then to retail or mail-order pharmacy, and then to consumer.¹⁰⁸ The obvious reason for such a system is to distribute the drug as widely as possible, which naturally increases revenues by facilitating consumer access. Limited distribution schemes, in contrast, eliminate the wholesaler and involve distribution only through specialty pharmacies selected by the manufacturer.

Such a restriction entails the brand's sacrifice of potential profits. Absent a medical reason to limit distribution (for example, monitoring patients), this restriction does not make business sense but can only be explained by its effect on generic rivals. The brand's refusal to sell the drug similarly would involve profit sacrifice. In fact, the sacrifice of profits itself provides a simple way to determine whether a company's sole motive is to impair competition. Such a sacrifice, which is economically irrational absent reduced competition, leads to the inference that the actor "was aware of and motivated solely to achieve that reduction."¹⁰⁹

In the regulatory context, and considering profit sacrifice, the cases discussed above foreshadow liability. For example, a generic that offers to purchase samples at the full retail price can claim that, under *Aspen Skiing*, the brand that refuses sales that would have been profitable was "willing to sacrifice short-run benefits and consumer goodwill in exchange for a perceived long-run impact on its smaller rival."¹¹⁰ Similar to the setting in *Otter Tail*, in which the defendant was able to "sell[] power at wholesale to those towns that wanted municipal plants" but refused to sell "solely to prevent municipal power systems from eroding its monopolistic position," the brand already is voluntarily selling the drug but restricting its distribution system so that it would not need to sell to others.¹¹¹ In addition, the *Trinko* Court's concerns are less relevant because the brand already sells

107. See *supra* text accompanying notes 12–21.

108. KAISER, FOLLOW THE PILL, *supra* note 21.

109. Steve Shadowen et al., *Anticompetitive Product Changes in the Pharmaceutical Industry*, 41 RUTGERS L.J. 1, 76 (2009); see also *supra* text accompanying notes 76–78.

110. *Aspen Skiing Co. v. Aspen Highlands Skiing Corp.*, 472 U.S. 585, 610–11 (1985).

111. *Otter Tail Power Co. v. United States*, 410 U.S. 366, 378 (1973); see Darren S. Tucker et al., *REMS: The Next Pharmaceutical Enforcement Priority?*, 28 ANTITRUST 74, 76 (2014).

at retail (reducing problems with “forced sharing”) and makes only a one-time sale (limiting judicial involvement).¹¹² At the same time, Turing’s change to the distribution scheme did not resemble the setting in *Trinko*, where “[t]he complaint d[id] not allege that Verizon voluntarily engaged in a course of dealing with its rivals,” but instead was similar to that in *Aspen Skiing*, where “[t]he unilateral termination of a voluntary (*and thus presumably profitable*) course of dealing suggested a willingness to forsake short-term profits to achieve an anticompetitive end.”¹¹³

The 2015 switch of pyrimethamine to a restricted distribution scheme as a condition of its sale to Turing could result in fewer sales and the sacrifice of profits. Turing left sales on the table by voluntarily cutting back its distribution scheme.¹¹⁴ Drug manufacturers typically have expansive distribution systems. Absent medical necessity, there is no reason to voluntarily restrict a distribution system, which would result in fewer sales. In this case in particular, there was no apparent reason to limit distribution 62 years after the FDA approved pyrimethamine and with no recent safety concerns. Turing would have no difficulty selling samples to any generic that requested them.¹¹⁵

If there were any doubt as to the reason for the change in the distribution system, it was dispelled by Turing itself. Jon Haas, the director of patient access at Turing, admitted that he “would block [a] purchase” of pyrimethamine if a generic manufacturer sought to order the pill and conceded that Turing “would like to do our best to avoid generic competition” and was “certainly not going to make it easier” for the generics.¹¹⁶ Turing’s insistence on behavior that lacks rational business sense provides strong evidence of blocking generic rivals. This is a powerful illustration of exclusionary conduct that violates the antitrust laws.

D. COUNTERARGUMENTS

There are four primary counterarguments that Turing did not engage in exclusionary conduct. First, there is no evidence that a generic has

112. *Verizon Commc’ns Inc. v. Law Offices of Curtis V. Trinko, LLP*, 540 U.S. 398, 408 (2004) (“Enforced sharing . . . requires antitrust courts to act as central planners, identifying the proper price, quantity, and other terms of dealing—a role for which they are ill suited. Moreover, compelling negotiation between competitors may facilitate the supreme evil of antitrust: collusion.”); *see* Tucker et al., *supra* note 111, at 76.

113. *Trinko*, 540 U.S. at 409.

114. *See supra* text accompanying notes 21–25.

115. *See* HOW TO OBTAIN A LETTER FROM FDA, *supra* note 27, at 2.

116. Ed Silverman, *How Martin Shkreli Prevents Generic Versions of his Pricey Pill*, PHARMALOT (Oct. 5, 2015), <http://pharmalot.com/how-martin-shkreli-prevents-generic-versions-of-his-pricey-pill/> [https://perma.cc/4BWA-8AAF].

attempted to obtain samples. Second, a price increase does not automatically demonstrate monopolization. Third, a company should not be forced to share its product with rivals. Fourth, courts typically treat exclusive distributor agreements as procompetitive.

First, there is no available evidence that a generic has attempted to obtain samples from Turing. A monopolization claim based on a refusal to license typically includes a request for a license, so some might assert that the absence of a refusal precludes an antitrust claim. But monopolization case law makes clear that there need not always be a formal request and refusal.

The Supreme Court has explained that plaintiffs may be able to show causation if making a request would be futile. For example, in *Zenith Radio Corp. v. Hazeltine Research, Inc.*, the Court made clear that a company's "fail[ure] to make a formal request for a [patent] license . . . can properly be attributed to [its] recognition that such a request would have been futile," as the defendant "had made its position entirely clear, and under these circumstances the absence of a formal request is not fatal to [the plaintiff's] case."¹¹⁷ In *Continental Ore v. Union Carbide & Carbon*, the Court did "not believe that [defendants'] liability under the antitrust laws can be measured by any rigid or mechanical formula requiring [the plaintiff] both to demand materials from respondents and to exhaust all other sources of supply."¹¹⁸ And in *Hanover Shoe v. United Shoe Machinery*, the Court "agree[d] with the courts below that in the circumstances of this case it was unnecessary for [the plaintiff] to prove an explicit demand" to purchase the defendant's machines.¹¹⁹

Other courts have similarly applied the futility rule. For example, in *Sullivan v. NFL*, the First Circuit held that a team owner did not need to "call for a vote and obtain an official refusal from the NFL" on its public-ownership policy since "such a request would be futile."¹²⁰ In *Chicago Ridge Theatre Limited Partnership v. M & R Amusement Corporation*, the Seventh Circuit explained that when the defendant's policy of providing films was well known, "the formality of the [plaintiff's] demands or bids . .

117. 395 U.S. 100, 120 n.15 (1969).

118. 370 U.S. 690, 699 (1962).

119. 392 U.S. 481, 488 (1968).

120. 34 F.3d 1091, 1104 (1st Cir. 1994); *see also id.* ("[O]fficial request and official refusal is not necessary to establish causality . . . [where] there is other evidence showing that defendant's practice caused injury in fact to the plaintiff.").

. cannot be a decisive issue in light of the futility of the requests.”¹²¹ And in *Out Front Productions v. Magid*, the Third Circuit explained that “[a]ntitrust suits are subject to no prerequisite, such as that imposed . . . for shareholder derivative suits, requiring that the complaint allege a demand or show futility.”¹²²

In this case, generic firms could reasonably argue that making a request would have been futile. The director of patient access at Turing conceded that he “would block [a] purchase” of pyrimethamine if a generic firm sought to order the pill, and conceded that Turing “would like to do our best to avoid generic competition.”¹²³ Such evidence supports a claim of futility.¹²⁴

Even beyond the futility claim, a generic might be able to show harm based on its ability, incentive, and preparation to enter the market. Though no company has yet announced an intention to enter the U.S. market, in India alone more than 40 companies manufacture generic versions of pyrimethamine.¹²⁵ These firms include large manufacturers with an international presence, such as Lupin Laboratories and Torrent Pharmaceutical, which could be attracted by the increased price given that they sell the product domestically for as little as \$0.03 per tablet.¹²⁶ It would be straightforward for these firms to request a sample and begin preparations to enter the U.S. market. In fact, it is even possible that some

121. 855 F.2d 465, 470 (7th Cir. 1988); *see also id.* (deeming conclusion “consistent with the general rule that a rigid demand requirement is not appropriate in antitrust cases”).

122. 748 F.2d 166, 169 (3d Cir. 1984); *see also id.* (“Treating a ‘demand’ by an antitrust plaintiff as if it were a condition precedent to maintenance of the suit misdirects the relevant focus, which should be on whether plaintiff has adduced the requisite proof of causation [N]o persuasive reason has been suggested why evidence of a demand is the only way to prove causation.”).

123. *Id.* Such a response is similar to that in *Aspen Skiing Co. v. Aspen Highlands Skiing Corp.* See 472 U.S. 585, 594 n.14 (1985) (“When the Highlands official inquired why Ski Co. was taking this position considering that Highlands was willing to pay full retail value for the daily lift tickets, the Ski Co. official answered tersely: ‘we will not support our competition.’”).

124. Alternatively, a generic could make a request and be turned down. *See* European Commission, *Microsoft Case*, <http://ec.europa.eu/competition/sectors/ICT/microsoft/investigation.html> [<https://perma.cc/6TWJ-QERS>] (last visited Feb. 18, 2016) (European Commission case against Microsoft originated with complaint from Sun Microsystems alleging that Microsoft refused to supply it with necessary information to interoperate with Microsoft’s dominant PC operating system).

125. Samir K. Brahmachari & Nisha Chandran, *Affordable Healthcare Threatened! Concern for All Stakeholders*, 109 CURRENT SCIENCE 1375, 1375 (2015).

126. *Id.*

companies may have recently requested samples without publicizing such a request.¹²⁷

Second, a price increase may not constitute monopolization under U.S. law. Such an argument contends that U.S. courts do not regulate price and that antitrust law is ill equipped to referee these disputes. The critique falls short, however, because the behavior targeted by the antitrust analysis is not the price increase but the restriction of the distribution system.

Turing's price increase is useful in revealing monopoly power.¹²⁸ If Turing lacked such power, it would not be able to impose and (in the face of extreme pressure) maintain a 5000 percent price increase. But the antitrust analysis in this section has targeted Turing's restriction of pyrimethamine's distribution system. As a natural result of drug firms' attempts to maximize profits, expansive networks are the typical distribution scheme in the drug industry. A company's restriction of its distribution network—especially after the drug has been on the market for 62 years and there are no new safety issues motivating the change—provides strong evidence that the conduct is exclusionary.

Third, a rebuttal asserts that a company should not be required to share its product with rivals. Antitrust law has famously declared that a company has the right “freely to exercise [its] own independent discretion as to parties with whom [it] will deal.”¹²⁹ But even that assertion often omits the crucial preface to the phrase: “In the absence of any purpose to create or maintain a monopoly.”¹³⁰

The case law makes clear that if a company undertakes actions that do not make sense unless they harm a rival, it typically will form the basis for liability.¹³¹ That is especially the case when the company makes a change in an existing, profitable practice.¹³² So when a company changes a

127. In addition to a claim by a generic against Turing, consumers could also challenge Turing's behavior on the grounds that it increased price and reduced output. Such effects would demonstrate consumers' antitrust injury as it would fall squarely within the range of injuries “of the type the antitrust laws were intended to prevent” and that “flow[] from that which makes defendants' acts unlawful.” *Brunswick Corp. v. Pueblo Bowl-O-Mat, Inc.*, 429 U.S. 477, 489 (1977).

128. As discussed above, the price increase also reflects the antitrust injury suffered by consumers. *See supra* note 127.

129. *United States v. Colgate & Co.*, 250 U.S. 300, 307 (1919).

130. *Id.*

131. *See Otter Tail Power Co. v. United States*, 410 U.S. 366 (1973).

132. *See Aspen Skiing Co. v. Aspen Highlands Skiing Corp.*, 472 U.S. 585, 608 (1985).

distribution network in place for decades with no apparent reason other than harming rivals, it should be subject to antitrust liability.

Fourth, Turing selected Walgreen's as an exclusive distributor.¹³³ Courts that have reviewed the practice of channeling distribution through a single dealer and refusing to sell to others have found such arrangements procompetitive because manufacturers generally have legitimate reasons for appointing exclusive distributors. For example, a distributor given sole rights to sell a manufacturer's product could be expected to use its best efforts to promote the product widely.¹³⁴

But Turing's relationship with Walgreen's is not a typical exclusive-distributor agreement. For starters, such arrangements tend to be employed by manufacturers that lack "interbrand" market power (in a market consisting of manufacturers selling different brands of the same type of product).¹³⁵ By contrast, and as discussed in detail above, Turing has not only market power but also monopoly power. As a result, it has the ability to injure competition by "deny[ing] . . . a needed or valuable input . . . to a rival."¹³⁶ Turing's refusal to provide samples to potential generic competitors harms the overall market as it increases price and reduces output in a way that an "intra-brand" restraint (within a single brand) does not. In fact, courts have held that antitrust liability could be warranted when a generic drug firm enters into an exclusive supply agreement to harm a rival.¹³⁷

In addition, unlike exclusive distribution agreements that involve "a combining of complements . . . for [the] greater good,"¹³⁸ Turing's distribution agreement does not offer any apparent efficiencies. There was no evidence that pyrimethamine was underused before the arrangement was

133. HOVENKAMP, *supra* note 31, ¶ 11.6d, at 654.

134. *E.g.*, Republic Tobacco Co. v. N. Atl. Trading Co., 381 F.3d 717, 736 (7th Cir. 2004); *see also* Planetarium Travel, Inc. v. Altour Int'l, Inc., 622 F. App'x 40, 41 (2d Cir. 2015) ("[E]xclusive distributorship arrangements are presumptively legal.").

135. *See* Leegin Creative Leather Products, Inc. v. PSKS, Inc., 551 U.S. 877, 890 (2007); HOVENKAMP, *supra* note 31, ¶ 11.6d, at 655.

136. A. Douglas Melamed, *Exclusionary Vertical Agreements, Address Before the ABA Antitrust Section*, U.S. DEPT. OF JUSTICE ANTITRUST DIVISION (Apr. 2, 1998), <http://www.justice.gov/atr/speech/exclusionary-vertical-agreements> [<https://perma.cc/FF73-ANQY>]; *see also* HOVENKAMP, *supra* note 31, ¶ 11.6d, at 655 (noting that when there is interbrand market power, "there may be cases where threats to competition are plausible").

137. *See* Geneva Pharms. Tech. Corp. v. Barr Labs. Inc., 386 F.3d 485, 504 (2d Cir. 2004) (exclusive supply agreement showed generic firm's "intent to seize the sole supply" of an active ingredient to harm a rival and "monopolize the generic [blood thinner] market").

138. Melamed, *supra* note 136.

implemented. And there were no apparent safety concerns¹³⁹ that justified the exclusive relationship. The timing of the change supports this conclusion, with the new system implemented for the first time 62 years after the drug entered the market.

V. ADDITIONAL EXAMPLES

The pyrimethamine example is not the only one raising antitrust concern based on restricted distribution. This Part presents additional issues raised by the monopoly power inquiry and then turns to other instances of exclusionary conduct.

First, as discussed above, a plaintiff must show monopoly power. A high market share, significant price increase, or output reduction could demonstrate monopoly power. This is especially the case when the behavior has received public scrutiny.¹⁴⁰ The inquiry, most generally, is whether the company has the ability to control prices and exclude competition. In making this determination, care must be taken to ensure that what initially appears to be a substitute is in fact a substitute. Given the high degree of importance that patients, physicians, and payers place on FDA approval in maintaining safety and potency for prescription drugs, a compounded drug cannot function as a large-scale substitute for an FDA-approved drug.

A similar argument can be made about prescription drugs imported from Canada or other countries. Though individual patients are permitted by the FDA to import drugs from Canada or other countries for their personal use under certain circumstances,¹⁴¹ such drugs are not widely viewed as legitimate substitutes for prescription drugs because they have not been approved by the FDA. Indeed, policymakers who seek to enhance this pathway as a way of improving patient access to lower-cost drugs inevitably design systems in which the FDA or another trusted regulatory authority certifies the reliability of a particular non-U.S.-based supplier first before they can sell their foreign products in the U.S. market.

139. See *supra* text accompanying notes 26–28.

140. See Austin Frakt, *Even Talking about Reducing Drug Prices Can Reduce Drug Prices*, N.Y. TIMES (Jan. 18, 2016), http://www.nytimes.com/2016/01/19/upshot/even-talking-about-reducing-drug-prices-can-reduce-drug-prices.html?smprod=nytcore-iphone&smid=nytcore-iphone-share&_r=0 [<https://perma.cc/Z6V7-XE57>] (stating that mere attention to issue of high drug prices resulted in drug companies lowering their prices).

141. *Is It Legal for Me to Personally Import Drugs?*, FDA (Dec. 28, 2015), <http://www.fda.gov/AboutFDA/Transparency/Basics/ucm194904.htm> [<https://perma.cc/NMY2-WFLN>].

Second, a plaintiff must show exclusionary conduct. The case law on monopolization sets the boundaries for such a determination. If a company makes a change to an existing profitable practice, that raises concern.¹⁴² So does the sacrifice of profits, which does not make sense absent its effect on competitors.¹⁴³ In particular, when a company restricts an existing, profitable distribution system without a pretense of promoting safety, careful scrutiny is warranted.

This analysis can be applied to other examples of restricted distribution schemes seemingly intended to forestall generic manufacturers. Two close precursors to the restricted distribution system in the pyrimethamine case arose with Shkreli's previous start-up company, Retrophin.

In 2014, Retrophin acquired chenodiol (Chenodal), another old, inexpensive drug used to treat a rare genetic disorder leading to deficiencies in cholesterol and bile acid breakdown that can cause neurologic dysfunction, cataracts, and cardiovascular disease.¹⁴⁴ Chenodiol was made available only through Retrophin's Chenodal Total Care Program that purports to assist patients with insurance needs, provides refilling and prescription delivery service, and offers adherence assistance to ensure that patients take medications as prescribed by their physicians.¹⁴⁵ When it established this program, Retrophin increased the price from \$9,460 to \$47,300 per 100 pills.¹⁴⁶ Chenodal now must be ordered over the phone from Retrophin's distribution partner, Dohmen Life Science Services. In fact, the company admitted that its "[c]losed distribution system does not allow for generics to access product for bioequivalence study."¹⁴⁷

142. *Aspen Skiing Co. v. Aspen Highlands Skiing Corp.*, 472 U.S. 585, 610–11 (1985) (“[T]he evidence supports an inference that Ski Co. was not motivated by efficiency concerns and that it was willing to sacrifice short-run benefits and consumer goodwill in exchange for a perceived long-run impact on its smaller rival.”).

143. *See id.*

144. *Retrophin Completes Acquisition of Manchester Pharmaceuticals*, BUSINESS WIRE (Mar. 27, 2014), <http://www.businesswire.com/news/home/20140327005607/en/Retrophin-Completes-Acquisition-Manchester-Pharmaceuticals> [<https://perma.cc/7HBG-KCGX>].

145. *Chenodal*, RETROPHIN, <http://www.retrophin.com/content/products/chenodal.php> [<https://perma.cc/A9J4-MZTM>].

146. *Retrophin, Inc. Report*, U.S. SEC. EXCH. COMM'N (Apr. 3, 2014), <http://ir.retrophin.com/secfiling.cfm?filingID=1193805-14-689&CIK=1438533> [<https://perma.cc/BB2M-9PUV>].

147. Derek Lowe, *The Most Unconscionable Drug Price Hike I Have yet Seen*, IN THE PIPELINE (Sept. 11, 2014), http://blogs.sciencemag.org/pipeline/archives/2014/09/11/the_most_unconscionable_drug_price_hike_i_have_yet_seen [<https://perma.cc/C4TS-E5PQ>].

Retrophin also owns the rights to tiopronin (Thiola), an old, inexpensive drug used to treat a rare condition called cystinuria, which predisposes patients to a certain type of kidney stone. Retrophin created a “Total Care Hub” program and raised the price from \$1.50 per pill to \$30 per pill (patients often require multiple pills per day).¹⁴⁸ After increasing the price, Retrophin stated that it would, “[s]imilar to Chenodal, . . . move Thiola into closed distribution.”¹⁴⁹ Thiola is distributed only to patients who fax enrollment paperwork directly to Retrophin and arrange for delivery. The company admitted that “[e]xclusivity (closed distribution) creates a barrier and pricing power.”¹⁵⁰ Similar, otherwise-unexplained behavior in these settings reveals a pattern of profit sacrifice with Shkreli’s companies, making even more apparent the concern with Turing’s conduct in the pyrimethamine case.

An example of a potentially problematic restricted distribution program not related to a Shkreli-led company was presented by the New York Attorney General’s lawsuit against Actavis (now Allergan) and its subsidiary Forest Laboratories (together Forest).¹⁵¹ As market exclusivity for its twice-daily Alzheimer’s disease medication, memantine (Namenda IR), was ending, Forest sought to introduce a once-daily extended-release version, memantine XR (Namenda XR).¹⁵² Forest first announced that it would stop distribution of memantine entirely in order to forcibly switch all memantine patients to memantine XR before generic memantine became available,¹⁵³ attracting the Attorney General’s attention for potentially illegally interfering with generic competition.

Forest then proposed an exclusive distribution contract with the mail-order specialty pharmacy Foundation Care, requiring all patients seeking memantine to receive the product through this intermediary and additionally requiring a special medical necessity form.¹⁵⁴ At the same time, the reformulated memantine XR would be made available through normal

148. *Id.*; see also Lydia Ramsey, *The CEO Who Jacked up the Price of a Drug by 5,000% Has Done This Before*, BUSINESS INSIDER (Sept. 23, 2015), <http://www.businessinsider.com/martin-shkreli-history-of-price-hikes-2015-9> [<https://perma.cc/2NP5-RGPA>].

149. Lowe, *supra* note 147.

150. Comm. Memorandum, *supra* note 38, at 3.

151. New York *ex rel.* Schneiderman v. Actavis PLC, 787 F.3d 638, 647–48 (2d Cir. 2015); see generally Vincent C. Capati & Aaron S. Kesselheim, *Drug Product Life-Cycle Management as Anticompetitive Behavior: The Case of Memantine*, 22 J. MANAGED CARE & SPECIALTY PHARM. 1 (2016).

152. *Actavis*, 787 F.3d at 647–48.

153. *See id.*

154. *See id.* at 648.

distribution channels.¹⁵⁵ The Attorney General challenged this proposal as well, ultimately securing a preliminary injunction that required Forest to continue the routine distribution of memantine until the generic versions of that product became available.¹⁵⁶ Restrictions of distribution systems that lack safety justifications and that are designed to restrict generic competition present conduct falling comfortably within the realm of exclusionary behavior that has been found to constitute monopolization.

VI. CONCLUSION

Across public discourse and the political system, Turing's significant price increase received significant attention. But the restriction of Turing's distribution system provides more of a hook for a potential antitrust claim. For starters, Turing appears to have monopoly power in engineering and maintaining a 5000 percent price increase, preventing hospitals from obtaining pyrimethamine, and ensuring the absence of FDA-approved substitutes for the drug.

Turing also appears to have engaged in exclusionary conduct in changing its distribution system in a way that sacrificed profits and only made sense in blocking generic competition. The combination of monopoly power and exclusionary conduct is the hallmark of a monopolization claim. Turing's behavior warrants close antitrust scrutiny.

155. Capati & Kesselheim, *supra* note 151.

156. *New York v. Actavis PLC*, No. 14 CIV. 7473, 2014 WL 7015198, at *43–46 (S.D.N.Y. Dec. 11, 2014), *aff'd sub nom.* *New York ex rel. Schneiderman v. Actavis PLC*, 787 F.3d 638 (2d Cir. 2015).

WIRELESS NETWORK NEUTRALITY: TECHNOLOGICAL CHALLENGES AND POLICY IMPLICATIONS

Christopher S. Yoo[†]

ABSTRACT

One key aspect of the debate over network neutrality has been whether and how network neutrality should apply to wireless networks. The existing commentary has focused on the economics of wireless network neutrality, but to date a detailed analysis of how the technical aspects of wireless networks affect the implementation of network neutrality has yet to appear in the literature. As an initial matter, bad handoffs, local congestion, and the physics of wave propagation make wireless broadband networks significantly less reliable than fixed broadband networks. These technical differences require the network to manage dropped packets and congestion in a way that contradicts some of the basic principles underlying the Internet. Wireless devices also tend to be more heterogeneous and more tightly integrated into the network than fixed-line devices, which can lead wireless networks to incorporate principles that differ from the traditional Internet architecture. Mobility also makes routing and security much harder to manage, and many of the solutions create inefficiencies. These differences underscore the need for a regulatory regime that permits that gives wireless networks the flexibility to deviate from the existing architecture in ways, even when those deviations exist in uneasy tension with network neutrality.

DOI: <https://dx.doi.org/10.15779/Z38HQ3RZ0S>

© 2016 Christopher S. Yoo.

[†] John H. Chestnut Professor of Law, Communication, and Computer & Information Science and Founding Director of the Center for Technology, Innovation and Competition, University of Pennsylvania. I would like to thank the New York Bar Foundation for its financial support. This paper benefited from presentations at ETH Zurich, University of Haifa, and the Free State Foundation. Responsibility for any errors remains with the author.

TABLE OF CONTENTS

I.	INTRODUCTION	1411
II.	THE FCC’S SPECIAL TREATMENT OF MOBILE BROADBAND	1414
A.	THE BASIC REGULATORY REGIME GOVERNING COMMUNICATIONS.....	1414
B.	THE 2010 OPEN INTERNET ORDER.....	1418
C.	THE 2015 OPEN INTERNET ORDER.....	1422
III.	THE BASIC ARCHITECTURAL COMMITMENTS UNDERLYING NETWORK NEUTRALITY	1424
A.	THE (SUPPOSED) ABSENCE OF PRIORITIZATION/QUALITY OF SERVICE	1424
B.	THE END-TO-END ARGUMENT.....	1427
1.	<i>The Absence of Per-Flow State</i>	1428
2.	<i>Unique, Universal Addresses Visible to All Other Machines</i>	1431
IV.	TRAFFIC GROWTH, BANDWIDTH CONSTRAINTS, AND NETWORK MANAGEMENT	1432
A.	DIFFERENCES IN WIRELINE AND WIRELESS QUALITY OF SERVICE AND RELIABILITY	1435
B.	DIFFERENT DIMENSIONS OF QUALITY OF SERVICE.....	1436
C.	CAUSES OF POOR QUALITY OF SERVICE ON WIRELESS BROADBAND NETWORKS	1437
1.	<i>Bad Handoffs</i>	1437
2.	<i>Local Congestion</i>	1438
3.	<i>The Physics of Wave Propagation</i>	1438
D.	IMPLICATIONS OF THE LOWER QUALITY OF SERVICE IN WIRELESS NETWORKS	1444
1.	<i>Error Correction</i>	1445
2.	<i>Congestion Management</i>	1446
E.	RESPONSES TO THE LOWER QUALITY OF SERVICE IN MOBILE BROADBAND NETWORKS.....	1447
V.	THE HETEROGENEITY OF DEVICES	1448
VI.	ROUTING	1449
A.	THE USE OF INTERNET GATEWAYS	1450
B.	ACCELERATION IN THE PACE OF CHANGES IN ROUTING ARCHITECTURE	1450
C.	COMPACTNESS OF THE ADDRESS SPACE	1451
D.	THE IDENTITY/LOCATOR SPLIT	1453
E.	MOBILE IP.....	1455
1.	<i>Security</i>	1457

2.	<i>Handoffs</i>	1457
3.	<i>Triangle Routing</i>	1457
VII.	CONCLUSION	1458

I. INTRODUCTION

For the past decade, a single issue dominated Internet policy debates: network neutrality. The perceived need to protect network neutrality led the Federal Communications Commission (FCC) to adopt its first Open Internet Order in 2010 only to see that order overturned on judicial review in 2014.¹ The FCC issued its second Open Internet Order in 2015, which was upheld by the courts the following year.² On April 27 2017, the FCC announced its agenda for its May 18 Open Meeting, which included a Notice of Proposed Rulemaking that would revisit most of the key provisions of the second Open Internet Order.³ The debate over network neutrality appears to be far from over.

Although myriad definitions of network neutrality exist,⁴ they share a general commitment to preventing network providers (such as Verizon and Comcast) that offer broadband access to end users from discriminating against traffic based on its source, destination, or content, or based on its associated application, service, or device. From this point of view, all application-specific intelligence and functionality should be confined to the computers operating at the edge of the network, while the routers operating in the core of the network should be kept as simple as possible.

1. Preserving the Open Internet, Report and Order, 25 FCC Rcd. 17905 (2010) [hereinafter 2010 Open Internet Order], *aff'd in part, vacated in part sub nom.* Verizon v. FCC, 740 F.3d 623 (D.C. Cir. 2014).

2. Protecting and Promoting the Open Internet, Report and Order on Remand, Declaratory Ruling, and Order, 30 FCC Rcd. 5601 (2015) [hereinafter 2015 Open Internet Order], *aff'd sub nom.* U.S. Telecom Ass'n v. FCC, 825 F.3d 674 (D.C. Cir. 2016).

3. Restoring Internet Freedom, Notice of Proposed Rulemaking, WC Docket No. 17-108 (FCC Apr. 27, 2017) [hereinafter 2017 Open Internet NPRM], https://apps.fcc.gov/edocs_public/attachmatch/DOC-344614A1.pdf.

4. See, e.g., Rachelle B. Chong, *The 31 Flavors of Net Neutrality: A Policymaker's View*, 12 INTELL. PROP. L. BULL. 147, 151–55 (2008) (identifying five distinct versions of network neutrality); Eli Noam, *A Third Way for Net Neutrality*, FIN. TIMES (Aug. 29, 2006, 5:26 PM), <http://www.ft.com/cms/s/2/acf14410-3776-11db-bc01-0000779e2340.html> (identifying seven distinct versions of network neutrality).

Designing the network in this manner is often regarded as essential to ensuring that the network remains open to all applications.⁵

One central issue in both Open Internet Orders was whether mobile broadband should be subject to less restrictive rules than fixed broadband. Specifically, the 2010 Open Internet Order adopted three rules, but restricted the application of one of the rules to mobile broadband and completely exempted mobile broadband from another rule.⁶ The 2015 Open Internet Order took a different approach, choosing to apply the same rules to both fixed and mobile broadband. At the same time, the 2015 Order repeatedly recognized the existence of key technical differences between fixed and mobile broadband that must be considered when determining whether a particular network management practice is permissible.⁷ The 2017 NPRM reopened this issue by “seek[ing] comment on whether mobile broadband should be treated differently from fixed broadband.”⁸

Both orders explicitly suggest that technical dissimilarities might justify the use of network management practices on mobile broadband networks that would not be allowed on fixed broadband networks. Indeed, the 2015 Order requires that regulators grapple with the technical details when determining whether a particular practice violates its terms. Unfortunately, the technical aspects of mobile broadband have gone largely unexplored. So far, the academic commentary has focused almost exclusively on the economics of wireless network neutrality, debating whether wireless broadband providers have the economic means and incentive to restrict traffic from certain sources or applications in ways

5. For the FCC’s most extensive elaboration of this rationale, see Preserving the Open Internet, Notice of Proposed Rulemaking, 24 FCC Rcd. 13,064, 13,070 ¶ 19, 13,086 ¶ 56, 13,088–89 ¶ 63 (2009) [hereinafter 2009 Open Internet NPRM]. For subsequent restatements embracing this principle, see 2015 Open Internet Order, *supra* note 2, at 5702 n.570; Protecting and Promoting the Open Internet, Notice of Proposed Rulemaking, 29 FCC Rcd. 5561, 5629 ¶ 8, 5597 ¶ 102 & n.226, 5702 n. 570, 5803 ¶ 431 (2014); Preserving the Open Internet, Broadband Industry Practices, Report and Order, 25 FCC Rcd. 17,905, 17,909–10 ¶ 13 & nn.13–14 (2010) [hereinafter 2010 Open Internet Order].

6. See 2010 Open Internet Order, *supra* note 5, at 17,956–62 ¶¶ 93–106; *Net Neutrality: Hearing Before the S. Comm. on Commerce, Sci. & Transp.*, 109th Cong. 9 (2006) (prepared statement of Vinton G. Cerf, Vice Pres. & Chief Internet Evangelist, Google Inc.) (“The remarkable success of the Internet can be traced to a few simple network principles—end-to-end design, layered architecture, and open standards . . .”).

7. 2015 Open Internet Order, *supra* note 2, at 5611 ¶ 34, 5643 ¶ 101, 5651 ¶ 118, 5665 ¶ 148, 5701 ¶ 218, 5703–04 ¶ 223.

8. 2017 Open Internet NPRM, *supra* note 3, at 30 ¶ 94.

that could harm consumers and innovation.⁹ While one can debate the economic merits of prohibiting discrimination and prioritization, to date the literature has not grappled with the technical challenges that wireless broadband providers face in managing their networks.

An examination of the way wireless broadband networks actually work at a technical level is thus essential to understanding how network neutrality should be applied to mobile broadband. As discussed further below, differences in the ways that wireless broadband networks manage congestion and reliability necessarily introduce far more intelligence into the core of the network than is the case with fixed broadband networks. Moreover, mobile broadband networks are subject to bandwidth constraints that are much more restrictive than those faced by fixed broadband networks, and mobile operators choose to mitigate congestion by treating traffic differently depending on the applications with which it is associated. Indeed, the engineering literature is replete with observations listing support for mobility as one of the key network functions that the current architecture fails to perform well.¹⁰ The National Science Foundation's Future Internet Architecture program is sponsoring a

9. The debate over how to apply network neutrality to mobile broadband networks was initiated by Tim Wu. See Tim Wu, *Wireless Carterfone*, 1 INT'L J. ON COMM. 389 (2007). For later discussions, see Babette E.L. Boliek, *Wireless Net Neutrality Regulation and the Problem with Pricing: An Empirical, Cautionary Tale*, 16 MICH. TELECOMM. TECH. L. REV. 1 (2009); George S. Ford, Thomas M. Koutsy, & Lawrence J. Spiwak, *A Policy and Economic Exploration of Wireless Carterfone Regulation*, 25 SANTA CLARA COMPUT. & HIGH TECH. L.J. 647 (2008); Rob Frieden, *Hold the Phone: Assessing the Rights of Wireless Handset Owners and Carriers*, 69 U. PITT. L. REV. 675 (2008); Robert W. Hahn, Robert E. Litan & Hal J. Singer, *The Economics of Wireless Net Neutrality*, 3 J. COMPETITION L. & ECON. 399 (2007); Gregory L. Rosston & Michael D. Topper, *An Antitrust Analysis of the Case for Wireless Network Neutrality*, 22 INFO. ECON. & POL'Y 103 (2010); Marius Schwartz & Federico Mini, *Hanging Up on Carterfone: The Economic Case Against Access Regulation in Mobile Wireless* (May 2, 2007) (unpublished manuscript), <http://ssrn.com/abstract=984240>.

10. See, e.g., Mark Handley, *Why the Internet Only Just Works*, 24 BT TECH. J. 119, 120 (2006); Raj Jain, *Internet 3.0: Ten Problems with Current Internet Architecture and Solutions for the Next Generation*, PROC. MIL. COMM. CONF. (MILCOM 2006) (2007), <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4086425>; Jon Crowcroft, *Net Neutrality: The Technical Side of the Debate*, ACM SIGCOMM COMPUTER COMM. REV., Jan. 2007, at 49, 51; Thrasylvoulos Spyropoulos et al., *Future Internet: Fundamentals and Measurement*, ACM SIGCOMM COMPUTER COMM. REV., Apr. 2007, at 101; Sixto Ortiz, Jr., *Internet Researchers Look to Wipe the Slate Clean*, COMPUTER, Jan. 2008, at 12.

project to explore how the Internet might need to be redesigned to accommodate mobile broadband.¹¹

Many of the ways that wireless broadband networks operate differently from fixed broadband networks involve explicit prioritization of certain types of applications. Other aspects of wireless broadband networks violate many central tenets of the Internet's architecture, either by changing the semantics or by changing the basic principles around which the Internet is currently designed. Such changes reduce the interoperability of the network and create a much tighter integration between end users and the network. Even less transformative proposals are likely to affect different applications and end users differently and inevitably cause traffic on wireless and wireline networks to behave in a strikingly different manner. Understanding the technical space is thus essential to understanding how and when differential regulatory treatment between wireline and wireless networks may be justified and determining how the exception for reasonable network management should be applied to wireless networks.

The balance of this Article is organized as follows: Part II lays out the relevant regulatory history. Part III explains the basic architectural principles generally associated with the Internet, specifically nondiscrimination and the end-to-end argument. Part IV describes the more restrictive bandwidth constraints that mobile broadband networks face. Part V discusses quality of service. Part VI examines the heterogeneity of devices, and Part VII addresses the additional complexity of routing.

II. THE FCC'S SPECIAL TREATMENT OF MOBILE BROADBAND

The FCC's attempts to mandate network neutrality have consistently recognized that mobile broadband faces greater challenges than fixed broadband. Indeed, these differences have led the FCC to apply fewer restrictions to mobile broadband than to fixed broadband.

A. THE BASIC REGULATORY REGIME GOVERNING COMMUNICATIONS

The basic structure of the laws governing U.S. communications technologies was established by the Communications Act of 1934 ("1934

11. MobilityFirst Future Internet Architecture Project, *Overview*, MOBILITYFIRST <http://mobilityfirst.winlab.rutgers.edu/> (last visited Feb. 9, 2016).

Act”) and has remained largely unchanged ever since.¹² Title II of the 1934 Act governs telecommunications services, which have historically consisted primarily of traditional telephone service provided via fixed-line technologies. Under Title II, telecommunications carriers are subject to common carriage regulation,¹³ which requires that they provide service to anyone who requests it on terms that are just, reasonable, and nondiscriminatory.¹⁴ A subsequent amendment to Title II authorizes the FCC to use a process known as “forbearance” to excuse telecommunications carriers from having to comply with any regulations that the FCC finds are not necessary to protect consumers.¹⁵

Title III of the 1934 Act governs spectrum-based communications, which initially consisted solely of radio and television broadcasting transmitted over the air. A provision of the 1934 Act prohibited broadcasting from being treated as common carriers.¹⁶ In *FCC v. Midwest Video Corp. (Midwest Video II)*, the Supreme Court held that this statutory provision prohibited the FCC from requiring any service regulated under the broadcasting statute from making its facilities available on a nondiscriminatory basis.¹⁷

The emergence of cellular telephony upset this tidy regulatory taxonomy by making it possible to provide telephone service via spectrum. In response, Congress amended Title III to permit regulating spectrum-based communications technologies as common carriers only if they constituted Commercial Mobile Services (CMS). A CMS is any mobile service that makes interconnected services available to the public.¹⁸ All other services are Private Mobile Services (PMS), which are exempt from common carriage regulation.¹⁹

The emergence of new services that combined the transmission associated with telephone service with the data processing and storage associated with modern computing raised the question of whether and how these technologies should be regulated. From the time these new services first emerged, the FCC consistently exempted them from most

12. Communications Act of 1934, Ch. 652, 48 Stat. 1064 (codified as amended at 47 U.S.C. §§ 151, 202, 212, 311, 313, 314, 316, 317, 506, 521, 543 (2012)).

13. 47 U.S.C. § 153(51) (1996).

14. 47 U.S.C. § 202(a) (1989).

15. 47 U.S.C. § 160(a) (1996).

16. 47 U.S.C. § 153(11) (2010).

17. 440 U.S. 689, 700–02, 707 (1979).

18. 47 U.S.C. § 332(c)(1)(A) (1996).

19. 47 U.S.C. § 332(c)(2) (1996).

regulation.²⁰ As then-FCC Chairman William Kennard could observe in 1999, “[f]or the past 30 years, the FCC has created a deregulatory environment in which the Internet could flourish.”²¹ That said, the FCC tried to avoid directly addressing the proper regulatory classification that would apply to broadband Internet access, which drew a sharp rebuke from two members of the Supreme Court in January 2002.²² Finally, in March 2002, the FCC ruled that cable modem service was not a Title II service.²³

The modern debate over network neutrality emerged in 2004, when a speech by FCC Chairman Michael Powell challenged the industry to preserve four “Internet freedoms.”²⁴ The first three freedoms called for allowing consumers to access legal content, run applications, and attach devices as they saw fit, while the fourth held that consumers should receive meaningful information about their service plans.²⁵

The Supreme Court’s *Brand X* decision eliminated any uncertainty about the propriety of the FCC’s 2002 decision regarding the regulatory classification of cable modem systems discussed above when it upheld the FCC’s ruling that the Internet was not a Title II service.²⁶ Although the Supreme Court noted that the FCC had not yet decided whether to impose

20. MTS and WATS Market Structure, Access Charge Reconsideration Order, 97 F.C.C.2d 682, 711–22 (1983).

21. William E. Kennard, Chairman, Fed. Commc’ns Comm’n, Remarks Before the Federal Communications Bar, Northern California Chapter: The Unregulation of the Internet: Laying a Competitive Course for the Future 2 (July 20, 1999), <https://transition.fcc.gov/Speeches/Kennard/spwek924.doc>.

22. Nat’l Cable & Telecomms. Ass’n v. Gulf Power Co., 534 U.S. 327, 348–49, 353–56 & n.5 (2002) (Thomas, J., joined by Souter, J., concurring in part and dissenting in part).

23. 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), *aff’d sub nom.* Nat’l Cable & Telecomm. Ass’n v. Brand X Internet Servs., 545 U.S. 967 (2005).

24. Michael K. Powell, Chairman, Fed. Commc’ns Comm’n, Remarks on Preserving Internet Freedom: Guiding Principles for the Industry Delivered at the Silicon Flatirons Symposium 5–6 (Feb. 8, 2004), http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-243556A1.pdf. For an earlier discussion, see Christopher S. Yoo, *Network Neutrality and the Economics of Congestion*, 94 GEO. L.J. 1847, 1857 (2006). The FCC considered the related issue of whether it should mandate open access to cable modem systems when clearing a series of cable industry mergers from 1999 to 2002. See Daniel F. Spulber & Christopher S. Yoo, *Access to Networks: Economic and Constitutional Connections*, 88 CORNELL L. REV. 885, 1015–18 (2003).

25. Powell, *supra* note 24, at 5.

26. Nat’l Cable & Telecomm. Ass’n v. Brand X Internet Servs., 545 U.S. 967 (2005).

any specific regulatory obligations on cable modem systems,²⁷ most observers believed that broadband Internet access services would not be subject to open access obligations.²⁸ Shortly thereafter, the FCC issued decisions ruling that broadband access provided by telephone companies and mobile providers were also not Title II services.²⁹

In 2005, the FCC issued a Policy Statement adopting four principles that echoed the four “Internet Freedoms” advanced in Powell’s speech.³⁰ The FCC’s first three principles mirrored Powell’s first three freedoms, albeit subject to some caveats.³¹ The FCC’s Policy Statement also replaced Powell’s transparency principle with an “entitle[ment] to competition among network providers, application and service providers, and content providers.”³²

The FCC was explicit that its Policy Statement was not a new set of rules. According to the FCC, the Policy Statement simply indicated its intention to “incorporate the above principles into its ongoing policymaking activities.”³³ FCC Chairman Kevin Martin released a concurrent statement recognizing that “policy statements do not establish rules nor are they enforceable documents” and expressing his confidence “that the marketplace will continue to ensure that these principles are maintained” and “therefore, that regulation is not, nor will be, required.”³⁴ Despite these concessions, the FCC invoked the Policy Statement as the basis for sanctioning Comcast for its use of Transmission Control Protocol

27. *Id.* at 996.

28. *See, e.g.*, John Blevins, *A Fragile Foundation — The Role of “Intermodal” and “Facilities-Based” Competition in Communications Policy*, 60 ALA. L. REV. 241, 279 n.155 (2009) (“In practice, . . . Title I ‘regulation’ is essentially deregulation.”).

29. *See* Appropriate Framework for Broadband Access to the Internet over Wireline Facilities, Report and Order and Notice of Proposed Rulemaking, 20 FCC Rcd. 14,853 (2005), *petition for review denied sub nom.* Time Warner Telecom, Inc. v. FCC, 507 F.3d 205 (3d Cir. 2007); Appropriate Regulatory Treatment for Broadband Access to the Internet over Wireless Networks, Declaratory Ruling, 22 FCC Rcd. 5901 (2007).

30. Policy Statement on Appropriate Framework for Broadband Access to the Internet over Wireline Facilities, Policy Statement, 20 FCC Rcd. 14,986, 14,988 (2005).

31. Specifically, the Policy Statement made the right to access applications “subject to the needs of law enforcement.” *Id.* It also limited the right to connect devices to “legal devices that do not harm the network.” *Id.* All of the principles were “subject to reasonable network management.” *Id.* at 14,988 n.15.

32. *Id.* at 14,988.

33. *Id.* at 14,988 & n.15.

34. Kevin J. Martin, Chairman, Fed. Comm’n Comm’n, Comments on Commission Policy Statement (Aug. 5, 2005), http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-260435A2.pdf.

(TCP) resets to slow down traffic generated by certain peer-to-peer file sharing applications in 2008.³⁵

Because wireless had not yet emerged as an important broadband platform, Chairman Powell's four freedoms and the 2005 Policy Statement did not draw any distinctions between different broadband technologies. The impetus to apply less restrictive network neutrality regulations to mobile broadband did not emerge until the proceedings that led to the 2010 Open Internet Order.

B. THE 2010 OPEN INTERNET ORDER

The first recognition that mobile broadband might receive separate treatment appeared in the 2009 Notice of Proposed Rulemaking (NPRM) that led to the 2010 Open Internet Order.³⁶ The NPRM proposed codifying the four principles included in the 2005 Policy Statement, augmented by new rules that prohibited discrimination and required transparency.³⁷ The NPRM also included an exception for reasonable network management.³⁸ The NPRM explicitly sought comment on whether nondiscrimination and reasonable network management might apply differently to mobile broadband.³⁹

The 2009 NPRM proved controversial from the outset.⁴⁰ It became even more so in April 2010, when the D.C. Circuit overturned the FCC's order sanctioning Comcast not because the FCC had failed to adopt formal rules, but rather because the FCC had failed to base its actions on any valid statutory grant of authority.⁴¹

Uncertainty about the FCC's jurisdiction over network neutrality led then FCC Chairman Julius Genachowski to float a proposal on May 6, 2010 that would have reversed the 2002 Cable Modem Declaratory Ruling upheld in *Brand X* and would have reclassified Internet access as a Title II service, thereby bringing the Internet within the regulatory regime that

35. Formal Complaint of Free Press and Public Knowledge Against Comcast Corporation for Secretly Degrading Peer-to-Peer Applications, Memorandum Opinion and Order, 23 FCC Rcd. 13,028, 13,050–58 ¶¶ 141–151 (2008), *rev'd sub nom.* Comcast Corp. v. FCC, 600 F.3d 642, 644 (D.C. Cir. 2010).

36. 2009 Open Internet NPRM, *supra* note 5.

37. *Id.* at 13,100–11 ¶¶ 88–132.

38. *Id.* at 13,112–15 ¶¶ 133–141.

39. *Id.* at 13,123–24 ¶¶ 171–174.

40. See Wendy Davis, *Controversy Continues as FCC Votes Unanimously to Consider Net Neutrality Rules*, MEDIA POST (Oct. 22, 2009, 6:21 PM), <http://www.mediapost.com/publications/article/115959/controversy-continues-as-fcc-votes-unanimously-to.html>.

41. See Comcast Corp. v. FCC, 600 F.3d 642, 644 (D.C. Cir. 2010).

governs traditional telephone service.⁴² Under the proposal, the FCC would also exercise its statutory forbearance authority to excuse broadband Internet access providers from all but six of the relevant statutory provisions.⁴³

Genachowski's reclassification proposal proved even more controversial than the 2009 NPRM. On May 24, 2010, seventy-four House Democrats signed a letter urging Genachowski not to reclassify broadband Internet access as a Title II service, warning that it would "jeopardize jobs" and "should not be done without additional direction from Congress."⁴⁴ Thirty-seven House Republicans filed a similar letter the same day.⁴⁵

Undeterred, the FCC adopted a Notice of Inquiry on June 17, 2010, seeking comment on the possibility of reclassifying the Internet as a Title II service, again over the objections of the two Republican Commissioners.⁴⁶ Over the summer, the FCC convened a series of closed-door meetings attempting to find common ground among the key industry players.⁴⁷ Concurrently, reports began to emerge that Google and Verizon were on the verge of announcing a new joint position on network neutrality.⁴⁸ Rumors of the impending agreement caused the FCC to end its efforts to broker a compromise.⁴⁹

Google and Verizon unveiled their joint proposal on August 10, 2010.⁵⁰ The joint proposal endorsed the FCC's vision of creating rules

42. Julius Genachowski, Chairman, Fed. Commc'ns Comm'n, *The Third Way: A Narrowly Tailored Broadband Framework 5–6* (May 6, 2010), https://apps.fcc.gov/edocs_public/attachmatch/DOC-297944A1.pdf.

43. *Id.*

44. Declan McCullagh, *Congress Rebukes FCC on Net Neutrality Rules*, CNET (May 24, 2010, 9:46 PM), <http://www.cnet.com/news/congress-rebuked-fcc-on-net-neutrality-rules/>.

45. *Id.*

46. *Framework for Broadband Internet Service, Notice of Inquiry*, 25 FCC Rcd. 7866, 7889–95 ¶¶ 52–66 (2010).

47. See Matthew Lasar, *A Peek Inside the "Secret, Backroom" Net Neutrality Meetings*, ARS TECHNICA (July 28, 2010, 3:56 PM), <http://arstechnica.com/tech-policy/2010/07/fcc-secret-net-neutrality-meetings-continue-in-plain-sight/>.

48. See Edward Wyatt, *Web Deal Near on Paying Up to Get Priority*, N.Y. TIMES, Aug. 5, 2010, at A1.

49. See *FCC Ends Net Neutrality Compromise Talks*, CBS NEWS (Aug. 5, 2010, 3:50 PM), <http://www.cbsnews.com/news/fcc-ends-net-neutrality-compromise-talks/>.

50. *Verizon-Google Legislative Framework Proposal*, GOOGLE BLOG (Aug. 10, 2010), http://www.google.com/googleblogs/pdfs/verizon_google_legislative_framework_proposal_081010.pdf; see also Alan Davidson & Tom Tauke, *A Joint Policy Proposal for an Open Internet*, GOOGLE PUB. POL'Y BLOG (Aug. 9, 2010), <http://googlepublicpolicy.blogspot.com/2010/08/joint-policy-proposal-for-open-internet.html>.

embodying the first three principles of the 2005 Policy Statement as well as the new rules mandating nondiscrimination and transparency.⁵¹ More importantly for our purposes, it provided a ringing endorsement of subjecting mobile broadband to less stringent regulation. Only the transparency principle would apply to mobile broadband “[b]ecause of the unique technical and operational characteristics of wireless networks, and the competitive and still-developing nature of wireless broadband services.”⁵² On September 1, 2010, the FCC issued a further inquiry seeking comment on how the proposed network neutrality rules should apply to mobile broadband in general and on the Google-Verizon joint proposal in particular.⁵³

The idea of subjecting mobile broadband to less stringent regulation than fixed broadband became embodied in the Open Internet Order that the FCC adopted on December 23, 2010.⁵⁴ Consistent with the Google-Verizon joint proposal, the 2010 Order applied the transparency rule to mobile broadband, but refrained from applying the nondiscrimination rule to mobile broadband.⁵⁵ The 2010 Order did part with the Google-Verizon joint proposal in one respect, by subjecting mobile broadband to a modified no-blocking rule applicable only to websites and “applications that compete with the provider’s voice or video telephony services.”⁵⁶ The rules were subject to exceptions for reasonable network management and specialized services.⁵⁷ Regarding legal authority, the 2010 Order opted not to regulate under Title II, and instead asserted a welter of other statutory provisions.⁵⁸

The FCC recognized that “mobile broadband presents special considerations that suggest differences in how and when open Internet protections should apply,” specifically that mobile broadband represented an early-stage platform characterized more competition and greater operational constraints.⁵⁹ Chairman Genachowski echoed this reasoning, noting that key differences distinguished mobile broadband from fixed broadband, including “unique technical issues involving spectrum and

51. *Verizon-Google Legislative Framework Proposal*, *supra* note 50, at 1.

52. *Id.* at 2.

53. Further Inquiry into Two Under-Developed Issues in the Open Internet Proceeding, Public Notice, 25 FCC Rcd. 12,637, 12,640–42 (2010).

54. 2010 Open Internet Order, *supra* note 5, at 17,958–62 ¶¶ 97–105.

55. *Id.* at 17,958 ¶ 96, 17,959 ¶ 98, 17,962 ¶ 104.

56. *Id.* at 17,959 ¶ 99.

57. *Id.* at 17,951–56 ¶¶ 80–92, 17,964–65 ¶¶ 112–114.

58. *Id.* at 17,966–81 ¶¶ 115–137.

59. *Id.* at 17,956–97 ¶¶ 94–95.

mobile networks, the stage and rate of innovation in mobile broadband; and market structure.”⁶⁰ The other two Democratic Commissioners expressed their wish that mobile broadband had been treated the same as fixed broadband, but nonetheless voted for the Order.⁶¹ Network neutrality advocates were not so easily satisfied, bringing a number of challenges to the decision to apply a lighter touch to mobile broadband.⁶²

The D.C. Circuit issued its decision resolving the various challenges to the 2010 Open Internet Order on January 14, 2014.⁶³ The court held that the FCC had the authority to regulate broadband Internet access under Section 706 of the Telecommunications Act of 1996,⁶⁴ but ruled that the FCC could not exercise that authority in a manner inconsistent with any other express statutory provisions, such as the section providing, “A telecommunications carrier shall be treated as a common carrier under this [Act] only to the extent that it is engaged in providing telecommunications services.”⁶⁵ Nondiscrimination is the hallmark of common carriage regulation (indeed the FCC explicitly equated its nondiscrimination rule with the nondiscrimination contained in Title II),⁶⁶ and the Communications Act prohibits the FCC from regulating any provider as a common carrier unless it were classified as a Title II provider—a step the FCC specifically declined to take with respect to broadband Internet access.⁶⁷ The court recognized that it had previously held that another access regulation mandating access on “commercially reasonable” terms did not constitute nondiscrimination because the regulation left “substantial room for individualized bargaining and discrimination in terms.”⁶⁸ The FCC’s reliance on the same rationale for both the nondiscrimination and the no-blocking rules led the court to strike down

60. *Id.* at 18,041.

61. *Id.* at 18,046 (Copps, Comm’r., concurring), 18,082 (Clyburn, Comm’r., approving in part and concurring in part).

62. *See, e.g.*, *Free Press v. FCC*, No. 11-2123 (1st Cir. filed Sept. 28, 2011); *Mountain Area Info. Network v. FCC*, No. 11-2036 (4th Cir. filed Sept. 27, 2011); *People’s Prod. House v. FCC*, No. 11-3905 ag (2d Cir. filed Sept. 26, 2011); *Media Mobilizing Project v. FCC*, No. 11-3627 (3d Cir. filed Sept. 26, 2011); *Access Humboldt v. FCC*, No. 11-72849 (9th Cir. filed Sept. 26, 2011). On October 6, 2011, the Judicial Panel on Multidistrict Litigation consolidated all of these appeals in the D.C. Circuit. *In re Fed. Commc’ns Comm’n, Preserving the Open Internet*, Report and Order, No. 1:11-ca-01356 (J.P.M.L. Oct. 6, 2011) (order granting motion to consolidate).

63. *Verizon v. FCC*, 740 F.3d 623 (D.C. Cir. 2014).

64. 47 U.S.C. § 1302 (2015).

65. *Verizon*, 740 F.3d at 635–50 (citing 47 U.S.C. § 153(51)).

66. *Id.* at 657.

67. *Id.* at 650–56.

68. *Id.* at 652 (quoting *Cellco P’ship v. FCC*, 700 F.3d 534, 548 (D.C. Cir. 2012)).

the no-blocking rule as well.⁶⁹ The court noted that the no-blocking rule might be reconstructed as a requirement of a minimum level of service, but found that argument barred by the FCC's failure to adopt such argument in the 2010 Order or to raise that argument in its briefs.⁷⁰

The D.C. Circuit also singled out the attempt to regulate mobile broadband for separate discussion. As noted above, a separate statutory provision provides that the FCC can only subject a mobile service to common carriage if it constitutes as a CMS, while barring common carriage regulation of PMS.⁷¹ Because the FCC had classified mobile broadband as a PMS, mobile broadband providers are statutorily immune from common carriage requirements "twice over."⁷² The invalidation of the no-blocking and the nondiscrimination rules rendered moot challenges to the decision not to apply them equally to both fixed and mobile broadband.

C. THE 2015 OPEN INTERNET ORDER

On May 15, 2014, four months after the D.C. Circuit's decision overturning the 2010 Open Internet Order's no-blocking and nondiscrimination rules, the FCC, now under the leadership of Chairman Tom Wheeler, issued a new Notice of Proposed Rulemaking designed to establish new rules.⁷³ The NPRM explicitly noted that it was following the "blueprint" laid out by the D.C. Circuit⁷⁴ by replacing the nondiscrimination rule with a mandate of commercial reasonableness.⁷⁵ It also adopted "the revised rationale the court suggested" and reconstructed the no-blocking rule to establish a minimal level of access.⁷⁶

The FCC tentatively decided to follow the approach taken by the 2010 Open Internet Order that subjected mobile broadband to a less stringent no-blocking rule and exempted mobile broadband from the nondiscrimination rule altogether, although the agency sought comment on whether it should revisit those decisions.⁷⁷ The FCC also sought

69. *Id.* at 658.

70. *Id.*

71. 47 U.S.C. § 332(c)(2) (1996).

72. *Verizon*, 740 F.3d at 650 (quoting *Cellco*, 700 F.3d at 538).

73. Protecting and Promoting the Open Internet, Notice of Proposed Rulemaking, 29 FCC Red. 5561 (2014).

74. *Id.* at 5563 ¶ 4, 5618 ¶ 162; *see also id.* at 5647 (statement of Chairman Tom Wheeler) (observing that the NPRM was designed to follow the "roadmap laid out by the court").

75. *Id.* at 5563 ¶ 3, 5594 ¶¶ 92–93, 5599–600 ¶¶ 110–111.

76. *Id.* at 5595 ¶ 95.

77. *Id.* at 5583–84 ¶ 62, 5598 ¶¶ 105–106, 5609 ¶ 140.

comment on whether it should continue to classify mobile broadband as a CMS and if so, whether forbearance should apply.⁷⁸

President Obama's November 20, 2014, endorsement of reclassifying broadband Internet access as a Title II service changed the course of the rulemaking dramatically.⁷⁹ Although Chairman Wheeler initially expressed some reluctance,⁸⁰ the Open Internet Order adopted by the FCC on February 26, 2015, explicitly reclassified broadband Internet access as a Title II service.⁸¹ Pursuant to this authority, the 2015 Order adopted three bright-line rules prohibiting blocking, throttling, and paid prioritization backed by a catch-all standard prohibiting unreasonable interference or disadvantage to consumers or edge providers.⁸² The blocking and throttling rules as well as the catch-all standard remained subject to reasonable network management.⁸³ The 2015 Order self-consciously revised the FCC's approach to mobile broadband.⁸⁴ In contrast to both the 2010 Order and the 2014 NPRM, the 2015 Order opted to apply the same rules to both fixed and mobile broadband.⁸⁵ Consistent with this change, the FCC reclassified mobile broadband as a CMS or its functional equivalent instead of a PMS.⁸⁶ The FCC continued to recognize that mobile networks "must address dynamic conditions that fixed, wired networks typically do not, such as the changing location of users as well as other factors affecting signal quality," as well as more restrictive capacity constraints.⁸⁷ The 2015 Order thus explicitly recognized that these challenges must be taken into account when assessing whether a practice constitutes reasonable network management and cautioned that this inquiry must preserve mobile broadband operators'

78. *Id.* at 5613–14 ¶ 149–150, 5616 ¶¶ 153, 155.

79. White House Office of the Press Secretary, Statement by the President on Net Neutrality (Nov. 10, 2014), <https://www.whitehouse.gov/the-press-office/2014/11/10/statement-president-net-neutrality>.

80. Brian Fung & Nancy Scuola, *Obama's Call for an Open Internet Puts Him at Odds with Regulators*, WASH. POST (Nov. 11, 2014), <https://www.washingtonpost.com/news/the-switch/wp/2014/11/11/the-fcc-weighs-breaking-with-obama-over-the-future-of-the-internet/>.

81. 2015 Open Internet Order, *supra* note 3, at 5618 ¶ 59, 5757–77 ¶¶ 355–87.

82. *Id.* at 5607–09 ¶ 14–22, 5609 ¶ 25, 5638 ¶ 92, 5685 ¶ 192.

83. *Id.* at 5699–704 ¶¶ 214–24.

84. *Id.* at 5635–43 ¶¶ 86–101.

85. *Id.* at 5609 ¶ 25, 5638 ¶ 92, 5685 ¶ 192. The FCC also sought comment on how the transparency rule should apply to mobile, *id.* at 5669 ¶ 155, and created a safe harbor for disclosures made in the format established by the Consumer Advisory Committee, *id.* at 5680 ¶ 179.

86. *Id.* at 5615 ¶ 48, 5776–90 ¶¶ 388–408.

87. *Id.* at 5703 ¶ 223 (footnotes omitted); *accord id.* at 5611 ¶ 34.

flexibility.⁸⁸ The D.C. Circuit upheld these aspects of the FCC's decision.⁸⁹ The 2017 NPRM sought comment on once again classifying mobile broadband as a CMS and reopened the question whether mobile broadband should be regulated differently from fixed broadband.⁹⁰

* * *

The FCC's network neutrality regulations have consistently acknowledged that the challenges associated with mobile broadband justify subjecting mobile broadband to lighter touch regulation than fixed broadband. In particular, the current rules require a detailed, context-specific assessment to determine whether a mobile operator's particular practice constitutes reasonable network management.

III. THE BASIC ARCHITECTURAL COMMITMENTS UNDERLYING NETWORK NEUTRALITY

The FCC ruled that mandating network neutrality was necessary to preserve two architectural features that have proven essential to promoting innovation.⁹¹ First, broadband Internet access providers had to be prevented from blocking or disadvantaging traffic associated with certain edge providers or applications.⁹² Second, regulators had to preserve the end-to-end architecture.⁹³ Each will be discussed in turn.

A. THE (SUPPOSED) ABSENCE OF PRIORITIZATION/QUALITY OF SERVICE

Network neutrality advocates often assert that the Internet is also based on the commitment not to permit routers to prioritize traffic based on its source, content, or the application with which it is associated.⁹⁴ Such prioritization would allow broadband providers to harm innovation by

88. *Id.* at 5611 ¶ 34, 5643 ¶ 101, 5651 ¶ 118, 5665 ¶ 148, 5701 ¶ 218, 5703–04 ¶ 223.

89. U.S. Telecom Ass'n v. FCC, 825 F.3d 674, 713–26 (D.C. Cir. 2016).

90. 2017 Open Internet NPRM, *supra* note 3, at 20–22 ¶¶ 55–62, 30 ¶ 94.

91. Some of these commitments fall outside the scope of this paper. One prime example is the idea of protocol layering. See Christopher S. Yoo, *Protocol Layering and Internet Policy*, 161 U. PA. L. REV. 1707 (2013). Another example is modularity. See Christopher S. Yoo, *Modularity Theory and Internet Policy*, 2016 U. ILL. L. REV. 1.

92. See 2010 Open Internet Order, *supra* note 5, at 17,915–23 ¶¶ 21–31.

93. See *id.* at 17,909–10 ¶ 13 & n.13.

94. See, e.g., 2010 Open Internet Order, *supra* note 5, at 17,947 ¶ 76 (“pay for priority would represent a significant departure from historical and current practice”); LAWRENCE LESSIG, *THE FUTURE OF IDEAS* 37 (2002) (arguing that “the design effects a neutral platform—neutral in the sense that the network owner can’t discriminate against some packets while favoring others”).

“preferring their own or affiliated content, demanding fees from edge providers, or placing technical barriers to reaching end users.”⁹⁵

As a matter of history, the claim that the Internet’s architecture did not permit prioritization is problematic.⁹⁶ Since its inception, the IP header has contained a six-bit *type of service field* designed to allow the attachment of different levels of priority to individual packets.⁹⁷ The original design accommodated three levels of precedence as well as additional flags for particular needs regarding delay, throughput, and reliability, although subsequent changes now allow this field to be used even more flexibly.⁹⁸

Moreover, claims that the Internet is hostile toward prioritization ignore certain realities about the routing architecture. Tier 1 ISPs share information about the routing architecture with one another through the Border Gateway Protocol (BGP). Enabling networks to engage in policy-based routing that alters the path that traffic takes based on its source or destination represented one of the principal motivations behind BGP’s most recent redesign.⁹⁹

Nor did efforts to support prioritization end there. Throughout the Internet’s history, the Internet Engineering Task Force (IETF) has issued standards designed to allow networks to provide differential levels of

95. See 2015 Open Internet Order, *supra* note 2, at 5629 ¶ 80.

96. See David D. Clark, *The Design Philosophy of the DARPA Internet Protocols*, ACM SIGCOMM COMPUTER COMM. REV., Aug. 1988, 108 (“The second goal [of the DARPA architecture after survivability] is that it should support . . . a variety of types of service. Different types of service are distinguished by differing requirements for such things as speed, latency and reliability.”); see also Kai Zhu, Note, *Bringing Neutrality to Net Neutrality*, 22 BERKELEY TECH. L.J. 615, 619–21, 634–38 (2007) (observing that the Internet was never designed to be neutral).

97. Info. Sci. Inst., *Internet Protocol: DARPA Internet Program Protocol Specification* 8, 18, 35–36 (Sept. 1981), <http://tools.ietf.org/pdf/rfc791> (IETF Network Working Group Request for Comments no. 791); see also Info. Sci. Inst., *DoD Standard Internet Protocol* 12, 26–27, (Dec. 1979), <http://128.9.160.29/ien/txt/ien123.txt> (Internet Engineering Note no. 123).

98. ANDREW S. TANENBAUM & DAVID J. WEATHERALL, *COMPUTER NETWORKS* 440 (5th ed. 2003).

99. CHRISTIAN HUITEMA, *ROUTING IN THE INTERNET* 195 (1995); Kirk Lougheed, A *Border Gateway Protocol (BGP)* 1 (June 1981), <http://tools.ietf.org/pdf/rfc1105> (IETF Network Working Group Request for Comments no. 1105). A leading textbook gives the following examples of policy-based routing: “1. Do not carry commercial traffic on an educational network. 2. Never send traffic from the Pentagon on a route through Iraq. 3. Use TeliSonera instead of Verizon because it is cheaper. 4. Don’t use AT&T in Australia because performance is poor. 5. Traffic starting or ending at Apple should not transit Google.” TANENBAUM & WEATHERALL, *supra* note 98, at 479.

quality of service, including Integrated Services (IntServ),¹⁰⁰ Differentiated Services (DiffServ),¹⁰¹ MultiProtocol Label Switching (MPLS),¹⁰² and such modern initiatives as Low Extra-Delay Batch Transport (LEDBAT).¹⁰³ Providing better support for quality of service (particularly for real-time data) was identified as one of the major goals of the transition to IPv6.¹⁰⁴ Indeed, IPv6 includes a *traffic class* field that is analogous to the type of service field in IPv4.¹⁰⁵ Moreover, IPv6 added a *flow label* field similar to the labels used by MPLS to incorporate prioritization and other routing policies.¹⁰⁶

To say that prioritization has a long historical pedigree is not to say that it has won the day. Just as quality of service has its advocates within the engineering community, it also has its detractors. If the presentations in the leading textbooks on network engineering are any guide, the controversy over quality of service shows no signs of abating, with people on both sides of the argument holding strong views.¹⁰⁷ This Article is not intended to take sides in this debate. Instead, the goal is simply to emphasize that the debate over the relative merits of prioritization remains far from settled. In any event, as the following discussion demonstrates, the arguments in favor of prioritizing certain applications over others becomes increasingly compelling when wireless networks are involved.

100. See Robert Braden et al., *Integrated Services in the Internet Architecture: An Overview* (June 1994), <https://tools.ietf.org/html/rfc1633> (IETF Network Working Group Request for Comments no. 1633).

101. See Steven Blake et al., *An Architecture for Differentiated Services* (Dec. 1998), <https://tools.ietf.org/html/rfc2475> (IETF Network Working Group Request for Comments no. 2475).

102. See Eric C. Rosen et al., *Multiprotocol Label Switching Architecture* (Jan. 2001), <https://tools.ietf.org/html/rfc3031>, (IETF Network Working Group Request for Comments no. 3031).

103. See Stanislav Shalunov et al., *Low Extra Delay Background Transport (LEDBAT)* (Dec. 2012), <https://tools.ietf.org/html/rfc6817> (IETF Network Working Group Request for Comments no. 6817).

104. Scott Bradner & Allison Mankin, *IP: Next Generation (IPng) White Paper Solicitation 4* (Dec. 1993), <http://tools.ietf.org/pdf/rfc1550> (IETF Network Working Group Request for Comments no. 1550); accord DOUGLAS E. COMER, INTERNETWORKING WITH TCP/IP 563 (5th ed. 2006); LARRY L. PETERSON & BRUCE S. DAVIE, COMPUTER NETWORKS: A SYSTEMS APPROACH 319 (4th ed. 2007); TANENBAUM & WEATHERALL, *supra* note 98, at 456.

105. Stephen E. Deering & Robert M. Hinden, *Internet Protocol, Version 6 (IPv6) Specification 25* (Dec. 1998), <http://tools.ietf.org/pdf/rfc2460> (IETF Network Working Group Request for Comments no. 2460).

106. *Id.*

107. See COMER, *supra* note 104, at 510, 515.

B. THE END-TO-END ARGUMENT

Another architectural principle often regarded as essential to enhancing innovation is known as the end-to-end argument.¹⁰⁸ In end-to-end system designs, the routers operating in the middle of the network are not optimized for any particular application; instead, any functionality needed to support particular applications is confined to the hosts operating at the edges of the network.¹⁰⁹ Restricting application-specific intelligence to the edges of the network allows developers of new applications to focus exclusively on the software running in the hosts and to avoid having to modify any application-specific programs running in the core of the network.¹¹⁰ This gives entrepreneurs the confidence that they will remain free to innovate without having to seek permission from any broadband Internet access providers.¹¹¹

Although end-to-end system design is sometimes treated as if it were an absolute mandate, it should instead be treated as a pragmatic rule of thumb that should give way under appropriate circumstances.¹¹² Even the IETF document that is most strongly associated with the principle recognizes that the continuous nature of technological change means that architecture principles inevitably change as well.¹¹³ This document observed that “[p]rinciples that seem sacred today will be deprecated tomorrow” and that “[t]he principle of constant change is perhaps the only principle of the Internet that should survive indefinitely.”¹¹⁴ As a result, the document rejected the idea that the end-to-end argument represented “dogma about how Internet protocols should be designed.”¹¹⁵ Indeed, the

108. The seminal statement of the end-to-end argument is found in J.H. Saltzer, D.P. Reed & D.D. Clark, *End-to-End Arguments in System Design*, 2 ACM TRANSACTIONS ON COMPUTING 277 (1984). For another leading statement, see Brian E. Carpenter, *Architectural Principles of the Internet* 2–3 (June 1996), <http://tools.ietf.org/pdf/rfc1958> (IETF Network Working Group Request for Comments no. 1958) [hereinafter RFC 1958].

109. 2009 Open Internet NPRM, *supra* note 5, at 13,070 ¶ 19.

110. *Id.*

111. *Id.* at 13,089 ¶ 63; *accord* Protecting and Promoting the Open Internet, Notice of Proposed Rulemaking, 29 FCC Rcd. 5561, 5564 ¶ 8; 2010 (2014); 2010 Open Internet Order, *supra* note 5, at 17,909–10 ¶ 13 & n.13.

112. Christopher S. Yoo, *Would Mandating Network Neutrality Help or Hurt Broadband Competition?: A Comment on the End-to-End Debate*, 3 J. ON TELECOMM. & HIGH TECH. L. 23 (2004).

113. RFC 1958, *supra* note 108, at 1.

114. *Id.*

115. *Id.* at 2.

document recognized that circumstances might cause the Internet Protocol to change altogether.¹¹⁶

The end-to-end argument is operationalized through two principles relevant to this Article. First, in an end-to-end design, routers do not maintain any information associated with any particular traffic. This is known as flow state or per-flow state.¹¹⁷ Second, each host should have a unique address that is visible to all other machines.¹¹⁸

1. *The Absence of Per-Flow State*

One of the central commitments around which the Internet is designed is that the routers operating in the core of the network store the individual segments comprising larger communication (known as packets) for the minimum time needed to forward them toward their final destination. As soon as the routers have finished forwarding the packets, the routers discard all information associated with them. Two corollaries of this principle are that each router makes its own decision about the direction to route any particular packet and that each packet travels through the network independent of the packets preceding or following it in the data stream. This concept represented a sharp change from the architecture around which the telephone network was designed, which established dedicated circuits between end users and channeled all of the data associated with that communication along that circuit. The switches in the core of such a circuit-switched network, such as the telephone network, must necessarily retain a lot of information about each flow passing through the network. This information about where packets came from or where they are routed to is called *per-flow state*.¹¹⁹

The Internet's origins as a military network meant that the architects placed the highest priority on *survivability*, measured by the network's continuing ability to operate despite the loss of nodes within the network.¹²⁰ Networks that rely on a large amount of per-flow state tend not to be particularly robust in this manner. Consider what occurs when a switch in the middle of a telephone network fails. When the switch is lost, so too is all of the information maintained by the switch with respect to each flow. The loss of this per-flow state means that neither the network

116. *Id.* at 3.

117. Clark, *supra* note 96, at 113 (flow state); Christopher S. Yoo, *The Changing Patterns of Internet Usage*, 63 FED. COMM. L.J. 7, 86 (2010) (per-flow state).

118. RFC 1958, *supra* note 108, at 5.

119. *Id.*

120. *See* Clark, *supra* note 96, at 106–07.

nor the end user can recover from this event. As a result, the communication fails, and the only way to reestablish it is by placing a new call. Designing the network to avoid per-flow state in the core of the network increased the network's survivability.¹²¹

That said, some entity involved in the communication must maintain per-flow state in order to monitor whether the communication was ever delivered. Should that entity fail the communication would necessarily fail as well. The Internet architects assigned responsibility for these function to the computers operated by end users at the edge of the network, called *hosts*, a practice that has become known as *fate sharing*. The rationale is that if the hosts involved in the communication fail, there is probably no need to finish the communication.¹²²

Although survivability represented the original justification for avoiding having routers operate in the core of the network to maintain per-flow state, this rationale has little applicability to the modern Internet. While the loss of nodes may be a real concern in the hostile environments in which the military operates, the destruction of nodes is not typically a major concern in commercial networks.¹²³ Instead, the modern rationale for avoiding the maintenance of per-flow state in the core of the network is to facilitate the interconnection of networks that operate on very different principles.

The manner in which the absence of per-flow state facilitates interconnection is well illustrated by the history of the Advanced Research Projects Agency Network (ARPANET), which is widely regarded as the predecessor to the Internet.¹²⁴ In the ARPANET, all of the routers operating in the core of the network, called Interface Message Processors or IMPs, were manufactured by a single company based on the same computer and ran the same software, and were interconnected by the same technology—telephone lines.¹²⁵ The IMPs were responsible for a wide variety of tasks. For example, consistent with the standard approach of day,¹²⁶ IMPs were responsible for making sure that the packets were

121. *Id.* at 108.

122. *Id.*; RFC 1958, *supra* note 108.

123. Clark, *supra* note 96, at 107.

124. See JANET ABBATE, *INVENTING THE INTERNET* 113–33 (1999).

125. F.E. Heart et al., *The Interface Message Processor for the ARPA Computer Network*, 36 AFIPS CONF. PROC. 551, 552 (1970).

126. See Geoff Huston, *The End of End to End?*, ISP COLUMN (May 2008), at 1, <http://www.potaroo.net/ispcol/2008-05/eoe2e.pdf> (noting that the predominant approach to digital networking during the 1970s and 1980s required that each switch in a path store

successfully delivered to the next IMP and, if not, for correcting any errors by resending the packets.¹²⁷ In addition, IMPs were responsible for congestion control.¹²⁸

As a result, IMPs had to maintain a large amount of information about the current status of the packets passing through its network. Although these tasks were often quite complex, the fact that all IMPs were constructed with the same technology and operated on the same principles made them very easy to interconnect. The architects encountered greater problems when they attempted to interconnect the ARPANET with the two other packet networks sponsored by the Defense Department: the San Francisco Bay Area Packet Radio Network (PRNET) and the Atlantic Packet Satellite Network (SATNET). Differences in transmission technologies, throughput rates, packet sizes, and error rates made these networks remarkably difficult to interconnect. In addition, every network would have to maintain the same per-flow state information as the other network with which it wanted to interconnect and would have to understand its expected response when receiving a communication from another router.¹²⁹

The International Network Working Group (INWG) considered a variety of solutions to these problems.¹³⁰ It rejected as too cumbersome and too error-prone approaches that would have required every host to run simultaneously every protocol used by other types of networks¹³¹ or would have required each system to translate the communication into another format whenever it crossed a boundary between autonomous systems as too cumbersome and error-prone.¹³² Instead, Vinton Cerf and Robert Kahn's seminal article creating the Internet Protocol (IP) established a single common language that all networks could understand.¹³³ To

a local copy of the data until it received confirmation that the downstream switch has received the data).

127. John M. McQuillan & David C. Walden, *The ARPANET Design Decisions*, 1 COMPUTER NETWORKS 243, 282 (1977).

128. Christopher S. Yoo, *Protocol Layering and Internet Policy*, 161 U. PA. L. REV. 1707, 1758 (2013).

129. See ABBATE, *supra* note 124.

130. *Id.* at 131–32.

131. Vinton G. Cerf & Robert E. Kahn, *A Protocol for Packet Network Interconnection*, 22 IEEE TRANSACTIONS ON COMM. 637, 638 (1974) (“The unacceptable alternative is for every HOST or process to implement every protocol . . . that may be needed to communicate with other networks.”).

132. See ABBATE, *supra* note 124, at 128; Vinton G. Cerf & Peter T. Kirstein, *Issues in Packet-Network Intercommunication*, 66 PROC. IEEE 1386, 1399 (1978).

133. Cerf & Kahn, *supra* note 131, at 638.

facilitate its use by multiple networks, this common language was kept as simple as possible and included only the minimum information needed to transmit the communication.¹³⁴ All of this information was placed in an internetwork header that every gateway could read without modifying it.¹³⁵ The fact that all of the information needed to route a packet was contained in the IP header eliminated the need for any router to know anything about the design of the upstream network delivering the packet to it or about the design of the downstream network to which it was delivering the packet.

This in turn meant that functions previously handled by routers, such as reliability, were now assigned to the hosts operating at the edge of the network. Even friendly observers have conceded that at the time this approach was regarded as “heresy,”¹³⁶ “unconventional,”¹³⁷ and “odd.”¹³⁸ Over time, it has become an accepted feature of the network.

2. *Unique, Universal Addresses Visible to All Other Machines*

The interconnection of different networks was further complicated by the fact that each network tended to employ its own idiosyncratic scheme for assigning addresses to individual hosts and routers.¹³⁹ The Internet’s architects solved this problem by requiring that all networks employ a single, uniform addressing scheme common to all networks.¹⁴⁰ This scheme included the address information in the header of every IP packet so that every router could access the address information directly instead of having to maintain per-flow state. Moreover, hosts operating at the edge of the network must make their IP addresses visible to the rest of the network.¹⁴¹

134. See Barry M. Leiner et al., *The DARPA Internet Protocol Suite*, IEEE COMM., Mar. 1985, at 29, 31 (“The decision on what to put into IP and what to leave out was made on the basis of the question ‘Do gateways need to know it?’”).

135. Cerf & Kahn, *supra* note 131, at 638–39.

136. Huston, *supra* note 126, at 1.

137. ABBATE, *supra* note 124, at 125.

138. Ed Krol & Ellen Hoffman, *FYI on “What Is the Internet?”* 2, 4 (May 1993), <http://tools.ietf.org/pdf/rfc1462> (IETF Network Working Group Request for Comments no. 1462).

139. See Cerf & Kahn, *supra* note 131, at 637.

140. See Cerf & Kirstein, *supra* note 132, at 1393, 1399 (discussing the common internal address structure required for packet-level interconnectivity); Cerf & Kahn, *supra* note 131, at 641 (“A uniform internetwork TCP address space, understood by each GATEWAY and TCP, is essential to routing and delivery of internetwork packets.”).

141. Tony Hain, *Architectural Implications of NAT* 7–8, 18 (Nov. 2000), <http://tools.ietf.org/pdf/rfc2993> (IETF Network Working Group Request for Comments no. 2993).

IV. TRAFFIC GROWTH, BANDWIDTH CONSTRAINTS, AND NETWORK MANAGEMENT

The sharp increase in bandwidth consumption poses one of the biggest challenges to wireless networks. Since 2010, the number of mobile broadband subscribers has exceeded the number of subscribers of all other broadband technologies combined.¹⁴² Moreover, industry observers estimate that wireless traffic will grow at an annual rate of 57% from 2014 to 2019, as compared with a growth rate of 23% forecast for fixed Internet service.¹⁴³ When traffic saturates the available capacity, packets are forced to wait in queues. These queues become sources of jitter and delay, which degrades the quality of service provided by the network.

The increase in the number of mobile broadband subscribers and the growth in wireless broadband traffic have increased the need for network providers to engage in network management. As a general matter, there are two classic approaches to managing explosive traffic growth. One solution is simply to increase network capacity.¹⁴⁴ The presence of additional headroom makes it less likely that spikes in traffic will saturate the network, which in turn allows the packets to pass through the network without any delay. The other solution employs network management to give a higher priority to traffic associated with those applications that are most sensitive to delay.¹⁴⁵

For example, traditional Internet applications, such as email and web browsing, are essentially file transfer applications. Because file transfer applications typically display their results only after the last packet is delivered, delays in the delivery of intermediate packets typically do not adversely affect their performance. This contrasts with real-time, interactive applications, such as Voice Over Internet Protocol (VoIP), video conferencing, and virtual worlds, which are becoming increasingly

142. Fed. Comm'n. Comm'n, Internet Access Services: Status as of December 31, 2013, https://transition.fcc.gov/Daily_Releases/Daily_Business/2014/db1016/DOC-329973A1.pdf.

143. See CISCO SYS., INC., CISCO VISUAL NETWORKING INDEX: FORECAST AND METHODOLOGY, 2014–2019, at 5 tbl.1 (2015), http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.pdf.

144. For a representative statement appearing in the engineering literature, see Yaqing Huang & Roch Guerin, *Does Over-Provisioning Become More or Less Efficient as Networks Grow Larger?*, PROC. 13TH IEEE INT'L CONF. ON NETWORK PROTOCOLS (ICNP) 225 (2005). For a similar statement appearing in the legal literature, see, for example, LESSIG, *supra* note 94, at 47 (arguing in favor of addressing bandwidth scarcity by increasing capacity instead of implementing quality of service).

145. Christopher S. Yoo, *Beyond Network Neutrality*, 19 HARV. J.L. & TECH. 1, 21–23 (2005).

important on the Internet. The performance of these applications depends on the arrival time and spacing of every intermediate packet, with delays of as little as one third of a second being enough to render the service unusable.¹⁴⁶ As such, these applications are considerably more vulnerable to network congestion.¹⁴⁷

Networks can help protect the operation of time-sensitive applications either by expanding capacity or by giving their packets a higher priority. In the latter case, it is conceivable that the network need only to rearrange the order of the intermediate packets without affecting when the last packet will arrive. If so, network management can improve the performance of the time-sensitive application without having any adverse impact on the application that is less time sensitive. Even if small delays occur, with non-time-sensitive applications such as file-transfer, delays of a fraction of a second are virtually undetectable.

A review of leading computer networking textbooks reveals that the choice between these two approaches has long been a source of controversy in the engineering community with respect to wireline networks.¹⁴⁸ In the wireline context, engineering studies indicate that the amount of headroom needed to preserve quality of service without prioritization can be substantial.¹⁴⁹ Expanding bandwidth thus maintains simplicity, but requires the incurrence of significant capital costs. The additional cost associated with nonprioritized solutions increases the number of subscribers that a bandwidth expansion needs to breakeven, which in turn limits broadband deployment in ways that are likely to exacerbate the digital divide.¹⁵⁰ Network management, on the other hand, substitutes operating costs for capital costs, which allows them to be recovered as they are incurred. It does have the side effect of adding complexity to the network.

The tradeoff between these two approaches plays out much differently in the context of wireless networking. As an initial matter, wireless

146. International Telecommunication Union, ITU Recommendation G.114 (2003).

147. The problem is most acute for interactive video, such as video conferencing. For linear video (whether prerecorded or live), media players can ameliorate the jitter caused by congestion by delaying playback to buffer a quantity of packets so they may be released in a steady stream. Yoo, *supra* note 117, at 71.

148. See COMER, *supra* note 104, at 510, 515.

149. See M. Yuksel et al., *Quantifying Overprovisioning vs. Class-of-Service: Informing the Net Neutrality Debate*, PROC. 9TH INT'L CONF. ON COMPUTER COMM. & NETWORKS (2010), http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5560131.

150. Christopher S. Yoo, *Network Neutrality, Consumers, and Innovation*, 2008 U. CHI. LEGAL FORUM 179, 188, 229–32.

networks face limits that wireline networks do not face with regards to the number of end users that can be served in a particular area. A person connected to the Internet via a wireline technology (whether fiber, coaxial cable, or twisted pairs of copper) employs a signal that is narrowly channeled through space. This geographic limitation allows multiple end users to avoid interfering with one another even if they are sitting side by side.¹⁵¹

Wireless signals propagate quite differently. Unlike wireline signals, wireless signals propagate in an unchanneled manner in all directions.¹⁵² The signals of one user are thus perceived as noise by other end users. As Claude Shannon recognized in 1948, the increase in noise reduces the amount of usable bandwidth available to those other users.¹⁵³ The greater the density of users becomes, the more constricted the bandwidth becomes. This implies that there is an absolute limit to the density of end users who can use wireless broadband in any particular geographic area.¹⁵⁴

Even more importantly, wireless providers' options for expanding capacity are much more limited than for wireline networking. Wireless providers can increase bandwidth by deploying a larger number of microwave base stations operating at lower power or by deploying increasingly sophisticated receiving equipment. Such solutions are typically quite costly. Moreover, the gains from such strategies are finite. Once they are exhausted, the restrictions on the amount of spectrum allocated to any particular service sharply limits network providers' ability to expand capacity any further.¹⁵⁵

These bandwidth limitations require wireless networks to engage in extensive network management.¹⁵⁶ Specifically, if a subscriber in a low-

151. The fact that any electrical current creates some degree of radio frequency interference means that adjacent usage does create some interference. Any such interference occurs at very low power and can be minimized by proper shielding of the cables and the equipment.

152. Piyush Gupta & P.R. Kumar, *The Capacity of Wireless Networks*, 46 IEEE TRANSACTIONS ON INFO. THEORY 388 (2000).

153. Claude E. Shannon, *Communication in the Presence of Noise*, 37 PROC. INST. RADIO ENGINEERS 10 (1949).

154. Gupta & Kumar, *supra* note 152, at 391–92. Wireless operators can reduce this interference by using directional transmitters and receivers. Such solutions work only if you know the location of every sender and receiver. As such, they are poorly suited to wireless networking of mobile devices.

155. Charles Jackson et al., *Spread Spectrum Is Good—But It Does Not Obsolete NBC v. U.S.*, 58 FED. COMM. L.J. 245, 253–59 (2006).

156. See Charles L. Jackson, *Wireless Efficiency Versus Net Neutrality*, 63 FED. COMM. L.J. 445, 477 (2011).

bandwidth location is speaking on the telephone, the wireless network will prioritize the voice traffic and hold all email and other data traffic until the subscriber moves to a higher-bandwidth location or ends the call.¹⁵⁷ Other services rate-limit or prohibit video and other high-bandwidth services to ensure that a small number of users do not occupy all of the available bandwidth.¹⁵⁸ Technologies such as T-Mobile's Binge On adopt a different approach: it uses a strategy pioneered by MetroPCS¹⁵⁹ to reduce the bandwidth needed to convey video by reducing the resolution of all video to 480p. The bandwidth reductions are so significant that T-Mobile is able to exempt this traffic from counting towards end-users' data caps.¹⁶⁰ From a technical standpoint, this scheme does not work for non-video applications and thus cannot be employed in an application-agnostic way. A prohibition on prioritization based on applications would obstruct these types of network management tools from being deployed.

Prioritizing certain applications over others requires tight integration of the network and the device. The FCC noted as much when repealing the regulation barring network providers from bundling telecommunications services with the devices used by end-user, also known as customer premises equipment or CPE. The FCC recognized that the equipment that increasingly serves as enhancements to the network requires sophisticated interactions between the network and the device that was being impeded by the unbundling requirement.¹⁶¹ In other words, the device was part of the functionality of the network itself, a fact that renders calls for mandating that wireless broadband networks be open to all devices problematic.¹⁶²

A. DIFFERENCES IN WIRELINE AND WIRELESS QUALITY OF SERVICE AND RELIABILITY

Wireline and wireless broadband networks also differ in terms of their reliability. As anyone who has suffered through dropped calls on their mobile telephone recognizes, wireless technologies are much less reliable

157. Yoo, *supra* note 117, at 78.

158. *Id.* at 78–79.

159. Christopher S. Yoo, Wickard *for the Internet: Network Neutrality after Verizon v. FCC*, 66 FED. COMM. L.J. 415, 458–59 (2014).

160. Jason Cipriani, *T-Mobile's Binge On Streams 480p Video. Does It Matter?*, FORTUNE (Nov. 11, 2015, 12:55 PM), <http://fortune.com/2015/11/11/tmobile-480p-video/>.

161. Policy and Rules Concerning the Interstate, Interexchange Marketplace, Report and Order, 16 FCC Rcd. 7418, 7427 ¶ 16 (2001).

162. *See Wu, supra* note 9, at 395–401.

than wireline technologies. Part of the problem is the difficulty of seamlessly handing off a communication when a mobile wireless user transfers from one base station to another. Other problems are due to the physics of wave propagation, which can cause interference in wireless networks to arise in much more transient and unpredictable ways than in wireline networks.

These differences in reliability have implications for many basic architectural decisions for the Internet. For example, although the current network relies on hosts to correct errors by resending packets that are dropped, in a wireless world it is often more efficient to assign responsibility for those functions to routers operating in the core of the network. In addition, wireline networks rely on hosts to manage congestion on the Internet. For reasons discussed below, wireless networks' lack of reliability means that the traditional approach to congestion management will not work well on wireless. The result is that basic functions such as recovery from errors and managing congestion—two of the most fundamental functions performed by the network—operate far differently on wireless networks than on wireline networks.

B. DIFFERENT DIMENSIONS OF QUALITY OF SERVICE

The performance guarantees provided by different networking technologies, known as quality of service or QoS, can vary widely. Most commentators discuss quality of service in terms of guaranteed throughput rates. As a preliminary matter, it bears mentioning that the engineering community typically views quality of service as occupying more dimensions than mere bandwidth. In addition, networks vary in terms of their reliability (i.e., the accuracy with which they convey packets), delay or latency (i.e., the amount of time it takes for the application to begin functioning after the initial request is made), and jitter (i.e., variations in the regularity of the spacing between packets).¹⁶³

Interestingly, applications vary widely in the types of quality of service they demand. For example, the transfer of health records is not particularly bandwidth intensive and can accept millisecond latencies and jitter without much trouble, but is particularly demanding in terms of reliability. Voice over Internet Protocol (VoIP) is also not bandwidth intensive and tolerates unreliability, but is quite sensitive to latency and jitter. Financial transactions have low bandwidth requirements, but must have latency guarantees in the microseconds and perfect reliability.

163. TANENBAUM, *supra* note 98, at 405.

Interactive video applications (such as video conferencing and virtual worlds) are bandwidth intensive and intolerant of jitter and latency, but can allow a degree of unreliability.

Furthermore, network systems can improve certain dimensions of quality of service, but only at the expense of degrading other dimensions.¹⁶⁴ For example, streaming video works best when packets arrive in a steady stream. As a result, it is quite sensitive to jitter. Irregularities in the spacing between packets can be largely eliminated by placing all of the arriving packets in a buffer for some length of time and beginning to release them later. The presence of an inventory of backlogged packets allows them to be released in a nice even pattern. The cost, however, is to create a delay before the application begins to run.

C. CAUSES OF POOR QUALITY OF SERVICE ON WIRELESS BROADBAND NETWORKS

Quality of service on wireless broadband networks can degrade for a wide variety of reasons not applicable to wireline networks. These reasons include bad handoffs between base stations, local congestion, and the physics of wave propagation.

1. *Bad Handoffs*

Bad handoffs represent an important cause of poor quality of service in mobile broadband networks. In order to receive service, a wireless device must typically establish contact with some base station located nearby. Circumstances may require a device to transfer its connection from one base station to another. For example, the mobile host may have moved too far away from the original base station. Alternatively, the current base station may have become congested or environmental factors may have caused the signal strength between the current base station and the mobile host to have deteriorated.¹⁶⁵ For reasons discussed more fully below, transferring responsibility for a mobile host from one base station to another has proven to be quite tricky. It is not unusual for wireless networks to make bad handoffs, which can cause communications to be dropped.

164. CHRISTOPHER S. YOO, *THE DYNAMIC INTERNET: HOW TECHNOLOGY, USERS, AND BUSINESSES ARE TRANSFORMING THE NETWORK* 25–27 (2012).

165. JAMES F. KUROSE & KEITH W. ROSS, *COMPUTER NETWORKING: A TOP-DOWN APPROACH* 572–74 (6th ed. 2013).

2. *Local Congestion*

In addition, because wireless technologies share bandwidth locally, they are more susceptible to local congestion than many fixed-line services, such as DSL and fiber to the home. Local congestion makes end users acutely sensitive to the downloading behavior of their immediate neighbors. Other technologies, such as cable modem systems, are also subject to local congestion. The more restrictive bandwidth limitations make this problem worse for wireless networks, as does the fact that wireless networks are typically designed so that data and voice traffic share bandwidth, unlike wireline telephone and cable modem systems which place their data traffic in a different channel from their core business offerings. As a result, wireless broadband networks are particularly susceptible to spikes in demand.

These limits have led many wireless providers to limit or ban bandwidth intensive applications, such as video and peer-to-peer downloads, in order to prevent a small number of users from rendering the service completely unusable. For example, some providers using unlicensed spectrum to offer wireless broadband in rural areas have indicated that they bar users from operating servers for this reason.¹⁶⁶ United blocks video on its airplanes. Amtrak similarly blocks video and restricts large downloads on its train, while permitting such traffic in its stations where bandwidth is less restricted.¹⁶⁷

3. *The Physics of Wave Propagation*

The unique features of waves can cause wireless technologies to face interference problems that are more complex and fast-changing than anything faced by wireline technologies. Anyone who has studied physics knows that waves have some distinctive characteristics. These characteristics can reinforce each other in unexpected ways, as demonstrated by unusual echoes audible in some locations in a room and by whispering corners, where the particular shape of the room allows sound to travel from one corner to the other even though a person speaks no louder than a whisper. As noise-reducing headphones and cars demonstrate, waves can also cancel each other out. Waves also vary in the extent to which they can bend around objects and pass through small

166. See, e.g., *Ensuring Competition on the Internet: Net Neutrality and Antitrust: Hearing Before the Subcomm. on Intellectual Prop., Competition, and the Internet of the H. Comm. on the Judiciary*, 112th Cong. 55 (2011) (prepared testimony of Laurence Brett (“Brett”) Glass, Owner and Founder, LARIAT).

167. Yoo, *supra* note 147, at 79 n.39.

openings, depending on their wavelength. The discussion that follows is necessarily simplified, but it is sufficient to convey the intuitions underlying some of the considerations that make wireless networking so complex.

For example, wireless signals attenuate much more rapidly with distance than do wireline signals, which makes bandwidth much more sensitive to small variations in how distant a particular user is from the nearest base station. This requires wireless providers to allocate bandwidth by dynamically requiring individual transmitters to adjust their power. The physics of wireless transmission can also create what is known as the “near-far” problem, where a transmitter can completely obscure the signal of another transmitter located directly behind it by broadcasting too loudly.¹⁶⁸ WiFi networks similarly adjust the power of individual users dynamically to help allocate bandwidth fairly.¹⁶⁹ Again, the solution is to require the nearer transmitter to reduce its power, and accordingly its available bandwidth, in order for the other transmitter to be heard.

Moreover, in contrast to wireline technologies, there is an absolute limit to the density of wireless users that can operate in any particular area. Shannon’s Law dictates that the maximum rate with which information can be transmitted given limited bandwidth is a function of the signal-to-noise ratio.¹⁷⁰ Unlike wireline transmissions, which travel in a narrow physical channel, wireless signals propagate in all directions and are perceived as noise by other receivers. That means that when more people use wireless broadband, the amount of bandwidth available to others operating in the same area is reduced. At some point, the noise becomes so significant that the addition of any additional wireless radios becomes infeasible.

Managing wireless networks is further complicated by the fact that waves are also subject to refraction and diffraction. Refraction is a change in speed and direction that occurs whenever the transmission medium through which the wave is passing changes, such as when a wave travelling through the air passes through a wall and then back into the air.

168. See, e.g., Mahesh K. Varanasi & Behnaam Aazhang, *Optimally Near-Far Multiuser Detection in Differentially Coherent Synchronous Channels*, 37 IEEE TRANSACTIONS ON INFO. THEORY 1006 (1991).

169. See, e.g., Huazhi Gong & JongWon Kim, *Dynamic Load Balancing Through Association Control of Mobile Users in WiFi Networks*, 54 IEEE TRANSACTIONS ON CONSUMER ELEC. 342 (2008).

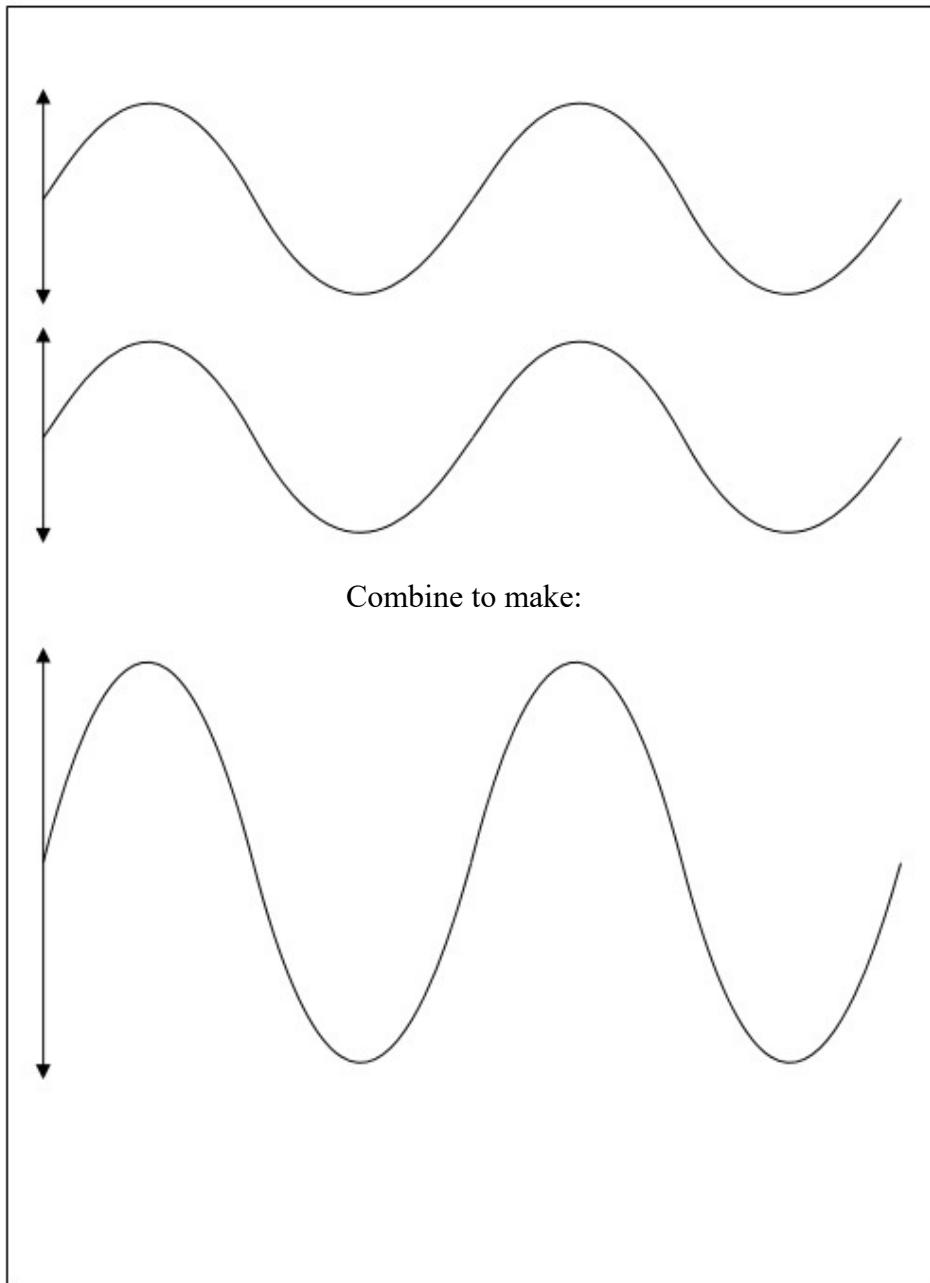
170. C. E. Shannon, *A Mathematical Theory of Communication* (pt. 1), 27 BELL SYS. TECH. J. 379 (1948); C. E. Shannon, *A Mathematical Theory of Communication* (pt. 2), 27 BELL SYS. TECH. J. 623 (1948).

The change in speed necessarily causes a change in the wave frequency. Diffraction occurs when a wave tries to bend around an obstacle or passes through a slit that is comparable in size to its wavelength. It has long been recognized that diffraction can cause complex patterns of interference.

Wireless transmissions also suffer from what are known as “multipath problems” resulting from the fact that terrain and other physical features can create reflections that can cause the same signal to arrive at the same location multiple times. Unless the receiver is able to detect that it is receiving the same signal multiple times, it will perceive multipathing as an increase in the noise floor that reduces the available bandwidth.¹⁷¹

When reflections cause the same signal to arrive by different paths, the signal can arrive either in phase (with the peaks and the valleys of the wave form from the same signal arriving at exactly the same time) or out of phase (with the peaks and the valleys of the wave form from the same signal arriving at different times). When waves reflecting off a hard surface arrive in phase, the signal reinforces itself, creating a localized hot spot in which signal is unusually strong.

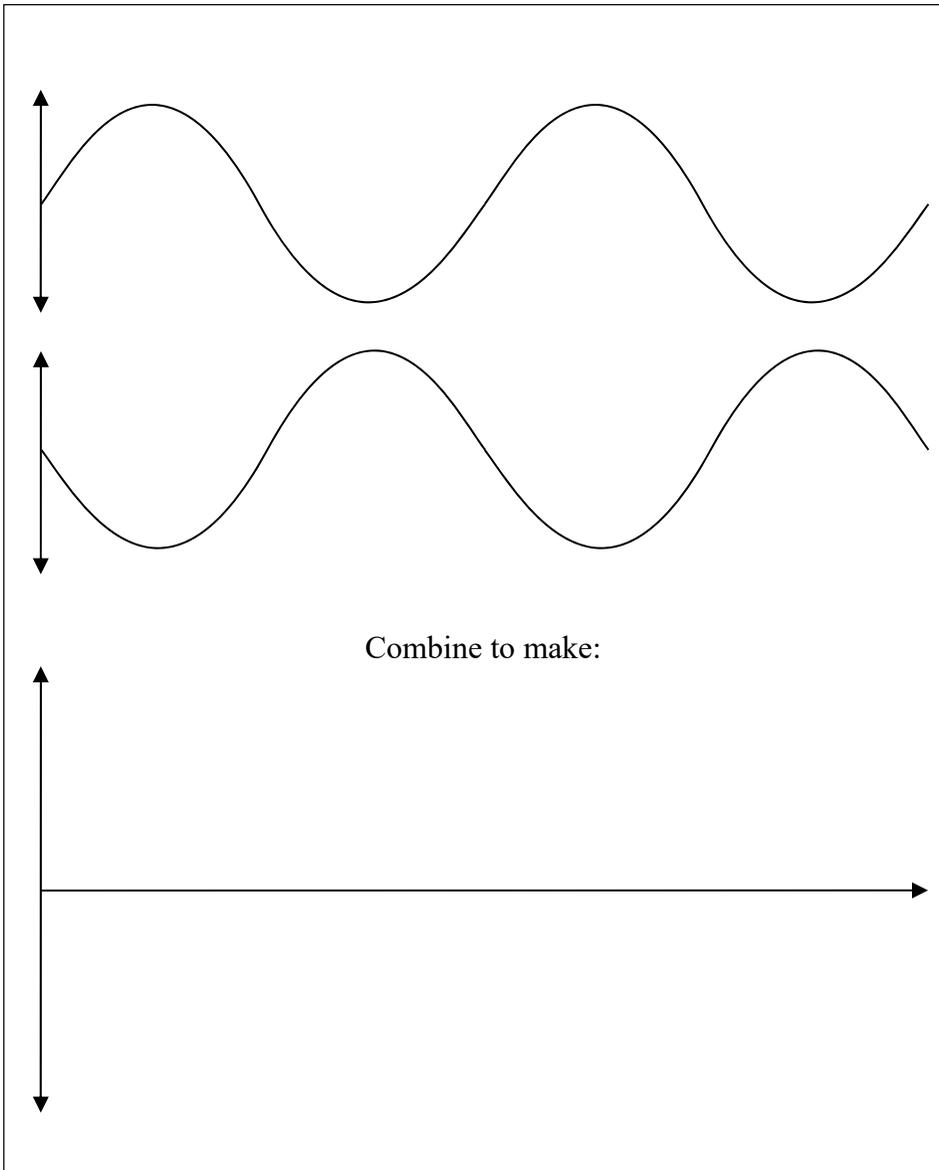
171. Jørgen Bach Andersen et al., *Propagation Measurement and Models for Wireless Communications Channels*, IEEE COMM., Jan. 1995, at 42.

Figure 1: Reinforcement of Two Wave Forms That Are in Phase

When reflected waves arrive out of phase, they can dampen the signal. When they arrive perfectly out of phase (i.e., 180° out of phase), the reflection can create a dead spot by canceling out the wave altogether. Although smart transmitters and receivers can avoid these problems if they know the exact location of each source and can even use the additional

signal to extend the usable transmission range, they cannot do so if the receiver or the other sources are mobile devices whose locations are constantly changing.

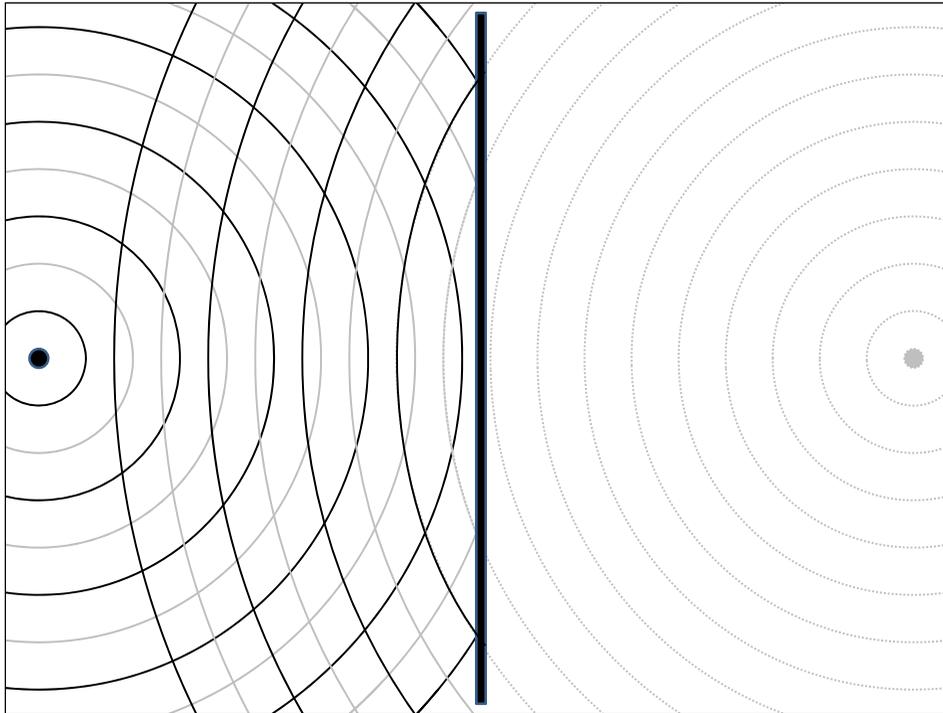
Figure 2: Cancellation by Two Wave Forms That Are 180° Out of Phase



A standard result in any physics textbook is that a reflection creates waves that are identical to a point source that is equidistantly located on the other side of the reflective surface and the signal strength is quite unpredictable. Consider the simple diagram in Figure 3, in which the black

circles represent the peaks of the wave form, while the grey circles represent the valleys. The points where two black circles or two grey circles cross represent hot spots where signals reinforce one another. The locations where a black circle crosses a grey circle represent dead spots where waves tend to cancel one another out.

Figure 3: Interference Caused by the Reflection of Waves



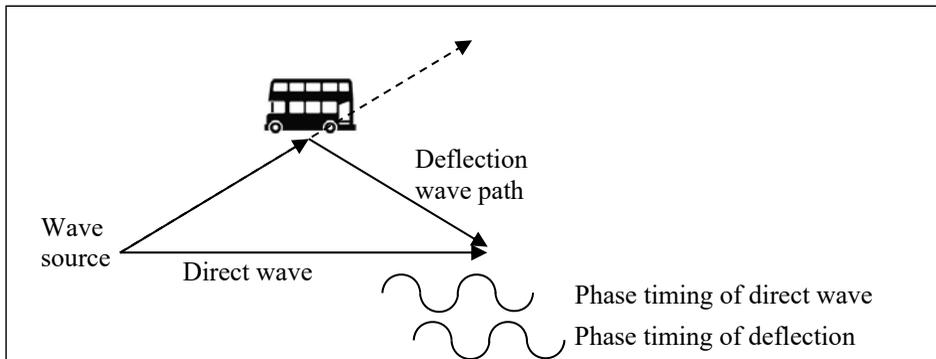
Obviously, individuals traversing a room might pass through a variety of hot and cold spots. In addition, wave reflections can result not only from immobile objects, such as terrain and buildings, but also from mobile objects, such as cars and trucks.¹⁷² The result is that the amount of bandwidth available can change dynamically on a minute-by-minute basis.

A participant at a May 2010 conference held at the University of Pennsylvania related a particularly vivid example of this phenomenon. While living in London, he had an apartment overlooking the famous Speakers' Corner in Hyde Park. Thinking that those in the Speakers' Corner might enjoy having WiFi service, he established a WiFi hotspot and pointed a directional antenna at the location only to find that his signal

172. *Id.*

was intermittently blocked even though nothing ever passed directly between his apartment and the Corner. He eventually discovered that the interference arose whenever a double-decker bus was forced to stop at a nearby traffic light. Even though the bus did not directly obstruct the waves travelling to and from the Speakers' Corner, it created a multipath reflection that periodically cancelled out the direct signal.¹⁷³

Figure 4: The Problem of Multipath Propagation



The result is that interference from other sources can be quite unpredictable and change rapidly from minute to minute. For these reasons, many wireless providers implement protocols that dynamically manage their networks based on the available bandwidth, giving priority to time-sensitive applications during times when subscribers are in areas of low bandwidth, such as by holding back e-mail while continuing to provide voice service. They have to implement these protocols much more aggressively and dynamically than do wireline providers.

D. IMPLICATIONS OF THE LOWER QUALITY OF SERVICE IN WIRELESS NETWORKS

The difference in the quality of service provided by wireless and wireline networks necessarily requires that the two networks be managed differently. In particular, wireless networks handle error correction and congestion in a manner that is quite different from wireline networks.

173. Christian Sandvig, Assoc. Professor of Commc'n, Univ. of Ill., Remarks presented at the Center for Technology, Innovation and Competition's conference on "Rough Consensus and Running Code: Integrating Engineering Principles into the Internet Policy Debates," *How to See Wireless* (May 7, 2010). For a description of the project, see PHILIP N. HOWARD, *NEW MEDIA CAMPAIGNS AND THE MANAGED CITIZEN* xi–xii (2006).

1. Error Correction

Wireless networks sometimes run afoul of the standard approach to ensuring reliability on the wireline Internet. The workhorse transport protocol on the Internet known as the Transmission Control Protocol or TCP ensures reliability by calling for every host to set a retransmission timer based on the expected round-trip time between the sending host and the receiving host.¹⁷⁴ Receiving hosts are supposed to send acknowledgements for every packet they successfully receive. If the sending host does not receive an acknowledgment when its retransmission timer expires, it resends the packet and repeats the process until it is successfully transmitted.¹⁷⁵

In many ways, relying on feedback loops and end-to-end retransmission is quite inefficient. Resending packets from the source requires the consumption of significant network resources. In addition, waiting for the retransmission timer to expire can cause significant delays. Such overhead costs become higher as the packet loss rates increase. If loss rates become sufficiently high, it may make sense for networks to employ network-based error recovery mechanisms instead of relying on end-to-end error recovery.

The lower reliability of wireless networks thus can lead system designers to deploy functionality in the core of the network to ensure reliability and error recovery. For example, PRNET employed a network-based reliability system known as forward-error correction.¹⁷⁶ The higher loss rates in wireless technologies also explains why wireless broadband networks are increasingly deploying network-based reliability systems, such as Automatic Repeat reQuest (ARQ), that detect transmission errors and retransmit the missing data from the core without waiting for the host-based retransmission timer to expire and without consuming the additional network resources needed to retrieve the packet all the way from the host.¹⁷⁷ Other techniques that allow routers in the core to participate in the transport layer exist as well.¹⁷⁸

174. TANENBAUM & WEATHERALL, *supra* note 98, at 569–70.

175. *Id.* at 568.

176. Robert E. Kahn et al., *Advances in Packet Radio Technology*, 66 PROC. IEEE, 1468, 1492 (1978).

177. KUROSE & ROSS, *supra* note 107, at 207–15; TANENBAUM & WEATHERALL, *supra* note 98, at 222–26.

178. *See* KUROSE & ROSS, *supra* note 107, at 575–77.

2. Congestion Management

The lack of reliability also requires that wireless technologies employ a significantly different approach to managing congestion. The primary mechanism for controlling congestion on the Internet was developed in the late 1980s shortly after the Internet underwent a series of congestion collapses. As noted earlier, TCP requires that receiving hosts send acknowledgments every time they successfully receive a packet. If the sending host does not receive an acknowledgement within the expected timeframe, it presumes that the packet was lost and resends it.¹⁷⁹ The problem is that the host now has sent twice the number of packets into a network that was already congested. Once those packets also failed to arrive, the host introduced still another duplicate packet. The resulting cascade would bring the network to a stop.

Because congestion is a network-level problem that is the function of what multiple end users are doing simultaneously rather than the actions of any one end user, some proposed addressing it through a network-level solution. This was done in the original ARPANET through networks running asynchronous transfer mode (ATM) and many other early corporate networks.¹⁸⁰ However, the router hardware of the time made network-based solutions prohibitively expensive. On the other hand, hosts can also stop congestion collapse if they cut their sending rates in half or more whenever they encounter congestion. The problem is that congestion is the product of what multiple hosts are doing, whereas any individual host only knows what it is doing. Thus the hosts operating at the edge of the network typically lack knowledge of when the network is congested.

Van Jacobson devised an ingenious mechanism by which hosts operating at the edge of the network can infer when the core of the network has become congested based on the information they were able to see. Jacobson noted that packet loss typically occurs for only two reasons: (1) transmission errors, or (2) discard by a router where congestion has caused its buffer to become full.¹⁸¹ Because wireline networks rarely drop packets due to transmission errors, hosts operating at the edge of the network could infer that the failure to receive an acknowledgement within

179. *Id.* at 240.

180. Raj Jain & K.K. Ramakrishnan, *Congestion Avoidance in Computer Networks with a Connectionless Network Layer: Concepts, Goals and Methodology*, PROC. COMPUTER NETWORKING SYMPOSIUM 134 (1988), <http://www.cse.wustl.edu/~jain/papers/ftp/cr1.pdf>.

181. Van Jacobson, *Congestion Avoidance and Control*, 18 ACM SIGCOMM COMPUT. & COMM. REV. 314, 319 (1988).

the expected time was a sign of congestion. Hosts could then take this as a signal to reduce congestion by slowing down their sending rates exponentially.¹⁸²

However, this inference is invalid for wireless networks. Wireless networks drop packets due to transmission error quite frequently, either because of a bad handoff as a mobile user changes cells or because of the interference problems discussed above. When a packet is dropped due to a transmission error, reducing the sending rate exponentially only serves to degrade network performance. Instead, the sending host should resend the dropped packet as quickly as possible without slowing down. In other words, the optimal response for wireless networks may well be the exact opposite of the optimal response for wireline networks.

E. RESPONSES TO THE LOWER QUALITY OF SERVICE IN MOBILE BROADBAND NETWORKS

In short, the deployment of wireless broadband is putting pressure on the traditional mechanisms for managing error correction and congestion, two of the most basic functions performed by the network. The higher loss rates make the traditional approach to error recovery more expensive and make it impossible to regard packet loss as a sign of congestion.

As a result, the engineering community is experimenting with a variety of alternative approaches.¹⁸³ One approach allows local recovery of bit errors through some type of forward error recovery.¹⁸⁴ One such solution places a “snoop module” at the base station that serves as the gateway used by wireless hosts to connect to the Internet that keeps copies of all packets that are transmitted and monitors acknowledgments passing in the other direction. When the base station detects that a packet has failed to reach a wireless host, it resends the packet locally instead of having the sending host do so.¹⁸⁵ A second approach calls for the sending host to be aware of when its transmission traverses wireless links. Dividing the transaction into two internally homogeneous sessions makes it easier to infer the current status of the network.¹⁸⁶ A third approach splits the

182. *Id.*

183. KUROSE & ROSS, *supra* note 107, at 576–77.

184. Ender Ayanoglu et al., *AIRMAIL: A Link-Layer Protocol for Wireless Networks*, 1 WIRELESS NETWORKS 47 (1995).

185. *See generally* Hari Balakrishnan et al., *Improving Reliable Transport and Handoff Performance in Cellular Wireless Networks*, 1 WIRELESS NETWORKS 469 (1995).

186. Ajay Bakre & B.R. Badrinath, *I-TCP: Indirect TCP for Mobile Hosts*, 1995 PROC. 15TH INT’L CONF. ON DISTRIBUTED COMPUTING SYS. (ICDCS ’95) 136, 137; Hari

wireless and the wireline approaches into separate TCP or UDP sessions.¹⁸⁷

Many of these approaches violate the semantics of TCP, since the packets are not addressed to the receiving hosts. Many of them introduce intelligence into the core of the network and violate the principle of avoiding per-flow state. The split connection approach violates the principle of end-to-end connectivity. All of them require introducing traffic management functions into the core of the network to a greater extent than originally envisioned by the Internet's designers.

V. THE HETEROGENEITY OF DEVICES

Starting with Michael Powell's 2004 four freedoms speech, every network neutrality proposal has called for broadband Internet access networks to be open to all legal devices. Indeed, the 2015 Open Internet Order included devices within the no blocking, no throttling, and no paid prioritization rules as well as the catchall prohibiting unreasonable interference and disadvantage.¹⁸⁸

In stark contrast to the fixed line world, wireless devices are not universally compatible with every network. For example, Verizon's wireless broadband network is based on a protocol known as Evolution-Data Optimized (EV-DO) operating in the traditional cellular portion of the spectrum. Sprint's wireless broadband network also employs EV-DO, but operates in the band of spectrum originally allocated to the second-generation wireless technology known as Personal Communications Services (PCS). AT&T's wireless broadband networks use a different format known as High Speed Packet Access (HSPA). Each has different technical characteristics. Indeed, the greater compatibility of HSPA with the iPhone is part of what led Apple initially to deploy the iPhone exclusively through AT&T.

Instead of relying on a personal computer, wireless broadband subscribers connect to the network through a wide variety of smart phones. These devices are much more sensitive to power consumption

Balakrishnan et al., *A Comparison of Mechanisms for Improving TCP Performance Over Wireless Links*, 5 IEEE/ACM TRANSACTIONS ON NETWORKING 756, 760 (1997).

187. Wei Wei et al., *Inference and Evaluation of Split-Connection Approaches in Cellular Data Networks*, PROC. ACTIVE & PASSIVE MEASUREMENT WORKSHOP (2006); Raj Yavatkar & Namrata Bhagwat, *Improving End-to-End Performance of TCP over Mobile Internetworks*, PROC. WORKSHOP ON MOBILE COMPUTING SYS. & APPLICATIONS 146, 147 (1994).

188. 2015 Open Internet Order, *supra* note 2, at 5607-09 ¶¶ 15-21.

than are PCs, which sometimes leads wireless network providers to disable certain functions that shorten battery life to unacceptable levels, for example because they either employ analog transmission or search constantly for an available connection. In addition, wireless devices have much less processing capacity and employ less robust operating systems than do the laptop and personal computers typically connected to wireline services. As a result, wireless devices are more sensitive to conflicts generated by multiple applications, which can cause providers to be much more careful about which applications to permit to run on them. This compels wireless broadband networks to manage devices and applications to a greater extent than wireline networks.

Wireless devices also tend to be much more heterogeneous in terms of operating systems and input interfaces including keyboards and touch screens. As a result, the dimensions and levels of functionality offered by particular wireless devices vary widely. It seems too early to predict with any confidence which platform or platforms will prevail. Furthermore, as noted earlier, many wireless networks address bandwidth scarcity by giving a higher priority to time-sensitive applications, which typically requires close integration between network and device. These features underscore the extent to which variations in particular devices are often an inextricable part of the functionality of the network.¹⁸⁹

These differences in compatibility and functionality call into question the provisions mandating that all broadband Internet access networks be open to all devices. Simply put, modern wireless devices prioritize traffic on the basis of application and are properly regarded as part of the network's functionality.

VI. ROUTING

Routing on wireless broadband networks is also very different from routing on fixed broadband networks. In particular, mobile broadband networks often exchange traffic with Internet gateways and rely on a legacy telephone technology to deliver traffic to end users instead of treating smartphones as IP-enabled devices. In addition, mobile broadband interferes with both the stability of routing tables and the compactness of the address space. Although potential solutions exist, such as the identity/locator split, they have yet to be implemented. As a result, wireless broadband networks must rely on a suite of protocols known as

189. Charles L. Jackson, *Wireless Efficiency versus Net Neutrality*, 63 FED. COMM. L.J. 445, 476–77 (2011).

mobile IP, which introduce a wide range of intelligence into the core of the network in ways that violate the end-to-end argument.

A. THE USE OF INTERNET GATEWAYS

One of the realities of wireless broadband networks is that they introduce a great deal of intelligence into the network in ways that fit less comfortably with the end-to-end argument. Recall that one of the Internet's foundational principles is that each host connected to the Internet must have a unique IP address that is visible and accessible to all other hosts. In addition, all of the routers within the network are supposed to route traffic on the basis of this address.

It bears mentioning that until recently, wireless networks have not routed traffic in this manner. Unlike devices connected to wireline networks, which have IP addresses that are visible to all other Internet-connected hosts, third-generation wireless devices did not have IP addresses. Instead, Internet connectivity is provided by an IP gateway located in the middle of the network that connects to individual wireless devices using a legacy telephone-based technology rather than IP. This means that for most of their history, wireless devices did not have the end-to-end visibility enjoyed by true Internet-enabled devices and instead connected through a virtual circuit between the Internet gateway and the wireless device. Fourth-generation wireless technologies such as LTE connect through IP. Until 3G is retired, some wireless devices will necessarily connect to the Internet on different and less open terms than devices connected through wireline networks.

This reality means that many wireless broadband devices violate the principle that each device has a unique IP address that is visible to all others. In addition, part of the connection operates using a different address system and employing circuit-based technologies that deviate from the Internet's commitment to store and forward routing. Simply put, traffic bound for and received from wireless devices will not pass through the network on the same terms as traffic going to and from hosts connected to the network through wireline technologies.

B. ACCELERATION IN THE PACE OF CHANGES IN ROUTING ARCHITECTURE

The mobility inherent in wireless broadband networks necessarily requires more frequent updates to routing tables than is the case for fixed broadband networks. Although solutions exist that could simplify this process, both the traditional version of the Internet Protocol, known as

IPv4, as well as the new version, known as IPv6, rely on a mobile IP approach that requires a great deal of intelligence in the network.

A key feature of the current routing architecture is that it is updated on a decentralized basis. Every backbone router periodically informs its adjacent neighbors of the best routes by which it can reach every location on the Internet. This means that initially any changes to the network architecture will only be advertised locally. During the next update cycle, routers that have been informed of the change will inform the routers located the next level away. Over time, the information will spread out in all directions until the entire network is aware of the change. When this occurs, the routing table is said to have reached equilibrium.

Before the routing table has reached equilibrium, however, some parts of the network may not know of certain changes that have occurred in other parts of the network. Suppose, for example, that one host in one corner of the network drops off the network. A host in a distant corner will not find out about that for quite some time. In the meantime, it could keep sending packets to a host that is no longer there, which wastes resources and unnecessarily adds to network congestion.

The efficient functioning of the network thus depends on the routing architecture being able to reach equilibrium. Whether it does so is largely a function of the speed with which locations change compared to the speed with which information about that change can propagate through the entire network. Moreover, the current architecture is built on the implicit assumption that Internet addresses change on a slower timescale than do communication sessions. So long as the address architecture changes at a slower timescale, any particular Internet-based communication may take the address architecture as given.

Mobility, however, increases the rate at which the address architecture changes. In addition, because addressing is handled on a decentralized basis, information about changes in the address architecture takes time to spread across the Internet. Increases in the rate with which the address space changes can cause communications sessions to fail and create the need for a new way to manage addresses.

C. COMPACTNESS OF THE ADDRESS SPACE

As a separate matter, wireless technologies are also causing pressure on the way the amount of resources that the network must spend on keeping track of Internet addresses. To understand why this is the case, one must keep in mind that routers typically follow one of two strategies in keeping routes. Some routers keep *global routing tables* that identify the outbound link that represents the most direct path to every single host

on the Internet. Other routers avoid the burden of maintaining complete routing tables by only keeping track of a limited number of paths. All traffic bound for locations for which this router does not maintain specific information is sent along a *default route* to a *default router*, which is responsible for identifying the route for delivery of all other traffic to its final destination.

The presence of default routes in a routing can give rise to a potential problem. For example, routers using default routes could point at one another, either directly or in a loop, which would cause the packets to pass back and forth indefinitely. The Internet ensures that traffic does not travel indefinitely through the network by assigning a *time to live* to each packet that limits the total number of hops that any packet may traverse before dropping off the network. Eventually, any packet caught in such a cycle will reach its maximum and drop off the network.¹⁹⁰

The best way to prevent such roads to nowhere is to ensure that at least some actors maintain global routing tables, which by definition are routing tables that do not include any default routes. This role is traditionally played by the major backbone providers, known as Tier 1 ISPs. More than the economic relationships (such as peering), many regard the maintenance of default free routing tables as the defining characteristic of Tier 1 ISPs.¹⁹¹

Sustaining a global routing table that maintained a separate entry for the best path to every location on the Internet has proved to be very difficult. The expansion of the Internet meant that the size of the routing table grew at a very fast rate. In fact, it grew faster than the routers could keep up.¹⁹²

The solution was an innovation called Classless InterDomain Routing (CIDR).¹⁹³ For our purposes, the important aspect of CIDR is that it allowed routers to use “route aggregation” to prevent routing tables from growing out of control. This mechanism can be illustrated by analogy to the telephone system. Consider an individual in Los Angeles who attempts to call the main telephone number for the University of Pennsylvania,

190. Paul Milgrom et al., *Competitive Effects of Internet Peering Policies*, in THE INTERNET UPHEAVAL 175, 179–80 (Ingo Vogelsang & Benjamin M. Compaine eds., 2000).

191. Peyman Faratin et al., *The Growing Complexity of Internet Interconnection*, 72 COMM. & STRATEGIES 51, 54 (2008).

192. Geoff Huston, *Analyzing the Internet BGP Routing Table*, 4 INTERNET PROTOCOL J., Mar. 2001, at 2, 3, <http://ipj.dreamhosters.com/wp-content/uploads/issues/2001/ipj04-1.pdf>.

193. Yoo, *supra* note 117, at 82.

which is (215) 898-5000. So long as all phones in the 215 area code are located in Philadelphia, a phone switch in Los Angeles could represent all of the telephone numbers in that area code ((215) xxx-xxxx) with a single entry in its routing table. Indeed, one can think of the millions of telephone numbers in the 215 area code as lying within the cone of telephone numbers represented by that entry.

Similarly, so long as all telephone numbers in the 898 directory within the 215 area code are connected to the same central office, switches within Philadelphia need not maintain separate entries for each phone number in that directory. Instead, they can represent the cone of all ten thousand telephone numbers located in (215) 898-xxxx with a single entry.

CIDR adopts a similar strategy to reduce the size of the routing tables maintained by Tier 1 ISPs. For example, the University of Pennsylvania has been assigned all of the addresses in the 128.91.xxx.xxx prefix (covering 128.91.0.0 to 128.91.255.255). Various locations have individual addresses falling within this range, with the main website for the University of Pennsylvania being covered by 128.91.34.233 and 128.91.34.234. Assuming that all of the hosts associated with these IP addresses are located in the same geographic area, a Tier 1 ISP could cover all of the one million addresses within this prefix with a single entry.

The success of this strategy depends on the address space remaining compact. In other words, this approach will fail if the 215 area code includes phone numbers that are not located in Philadelphia. If the telephones associated with those numbers sometimes lie outside the Philadelphia area, the telephone company will have to maintain separate entries in its call database for all phones located outside the area. Similarly, if some hosts with the 128.91.xxx.xxx prefix reside outside the Philadelphia area, Tier 1 ISPs will have to track those locations with additional entries in their routing tables.

The advent of mobile telephony and mobile computing means, of course, that telephones and laptops will often connect to the network outside their home locations. This in turn threatens to cause the routing tables to grow faster again. Other developments, including multihoming, the use of provider-independent addresses, and the deployment of IPv6, are further reducing the compactness of the routing table.

D. THE IDENTITY/LOCATOR SPLIT

A solution does exist that would not require introducing intelligence into other parts of the network to accommodate routing. This solution is

known as the identity/locator split.¹⁹⁴ The idea gained new impetus by the Report from the Internet Architecture Board (IAB) Workshop on Routing and Addressing, which reflected a consensus that such a split was necessary.¹⁹⁵ The International Telecommunication Union (ITU) has also embraced the need for the ID/locator split in Next Generation Networks (NGNs).¹⁹⁶ Additionally, it is the focus of a major research initiative sponsored by the National Science Foundation's Future Internet Architecture Program.¹⁹⁷

The proposal is based on the insight that an IP address currently plays two distinct functions. It simultaneously serves as an *identifier* that identifies a machine, and it serves a *locator* that identifies where that machine is currently attached to the network topology. When all hosts were connected to the Internet via fixed telephone lines, the fact that a single address combined both functions was not problematic. The advent of mobility caused the unity of identity and location to break down. A single mobile device may now connect to the network through any number of locations. Although the network could constantly update the routing table to reflect the host's current location, doing so would require propagating the updated information to every router in the network as well as an unacceptably large number of programs and databases.

Others have proposed radical changes in the addressing and routing architecture. One approach would replace the single address now employed in the network with two addresses: one to identify the particular machine and the other to identify its location.¹⁹⁸ Others criticize such proposals as unnecessarily complicated.¹⁹⁹

194. For an early statement, see Jerome H. Saltzer, *On the Naming and Binding of Network Destinations* (Aug. 1993), <http://tools.ietf.org/pdf/rfc1498> (IETF Network Working Group Request for Comments no. 1498) (identifying the potential need for separate names for nodes and network attachment points).

195. David Meyer et al., *Report from the IAB Workshop on Routing and Addressing* 22–23 (Sept. 2007), <http://tools.ietf.org/pdf/rfc4984> (IETF Network Working Group Request for Comments no. 4984).

196. INT'L TELECOMM. UNION, TELECOMM. STANDARDIZATION SECTOR, RECOMMENDATION ITU-T Y.2015: GENERAL REQUIREMENTS FOR ID/LOCATOR SEPARATION IN NGN (2009), http://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-Y.2015-200901-I!!PDF-E&type=items.

197. MobilityFirst Future Internet Architecture Project, *supra* note 11.

198. See Chakchai So-In, *Virtual ID: ID/Locator Split in a Mobile IP Environment for Mobility, Multihoming and Location Privacy for the Next Generation Wireless Networks*, 5 INT'L J. INTERNET PROTOCOL TECH. 142 (2010) (surveying alternative approaches to the ID/locator split).

199. See, e.g., Dave Thaler, Keynote Address at the 3rd ACM International Workshop on Mobility in the Evolving Internet Architecture (MobiArch 2008): Why Do

If deployed, the identity/locator split would represent a radical deviation from the existing architecture. Whatever solution is adopted would represent a fundamental change in the network layer than unifies the entire Internet. It would require a change in the way we approach routing and addressing and require reconfiguring every device attached to the network. If implemented, it would eliminate some of the asymmetries in the way that routing to mobile hosts is done. To date, however, the identity/locator split has not yet been implemented, and any future implementations would require an extended transition time during which networks would have to operate both modes.

E. MOBILE IP

Instead of relying on solutions that would have kept the network simple, the modern Internet relies on a complex system of protocols operating in the core of the network to accommodate mobility. The most straightforward approach to addressing mobility would be to assign a mobile host a new IP address whenever it changes location. This would put significant strain on the network by requiring that it inform the rest of the network about the change. To the extent that it disrupts the compactness of the address space, it may create additional pressure on the routing architecture by causing the routing table to grow. In addition, dynamically changing IP addresses in the middle of an application may cause many applications to fail.²⁰⁰

How, then, do we handle mobility without having to update the routing tables constantly and without causing the size of routing tables to grow out of control? The Internet currently solves these problems through a regime known as *mobile IP*. Under mobile IP, each mobile user has a *home network*, with all other networks labelled *foreign networks*. The mobile host designates a router located on its home network as the contact point for all IP-based communications directed to the mobile host. This contact point is called the *home agent*. Anyone seeking to contact the mobile host, called the *correspondent*, simply sends the packets to the home agent, which then forwards the communication to the mobile host. If the mobile host moves from one foreign network to another, it simply notifies its home agent, which then routes any new packets it receives to the new location.

We Really Want an ID/Locator Split Anyway? (Aug. 22, 2008), <http://conferences.sigcomm.org/sigcomm/2008/workshops/mobiarch/slides/thaler.pdf>.

200. PETERSON & DAVIE, *supra* note 104, at 290.

Although this solution sounds relatively simple, actually implementing it can be quite complex. For example, the home agent has to know to where the mobile host is currently located. This is relatively easy when the mobile host initiates the transaction. It is more complicated when a third party is attempting to contact the mobile host. Stated in the example of mobile telephony, networks can easily discover where a particular cellular user is located when it is that user that is initiating the call. The simple fact of establishing contact with the local microwave tower announces the location. The situation is different when the mobile user is receiving the call. To terminate this call, the network has to know where the mobile user is even when it is simply sitting around waiting.

This means that if a mobile host is to receive traffic, it must constantly announce its presence to the network serving its current location so that the network knows that it is there. This can be accomplished by designating a router located on the foreign network as the *foreign agent* responsible for managing mobile IP. Every mobile host must regularly register with the foreign agent serving the local foreign network in order to receive communications. This can happen by the foreign agent sending an advertisement notifying mobile nodes located in its service area that it is prepared to facilitate mobile IP or by the mobile node sending a solicitation to see if any foreign agents are located nearby capable of supporting mobile IP. Once a foreign agent registers the presence of a mobile host, it must then notify the home agent about the mobile host's current whereabouts so that the home agent knows where to forward any packets that it receives. Mobile IP works best if mobile nodes deregister when they leave the foreign network.

So how does the home agent send the packets to the foreign agent for delivery? It could alter the IP address contained in the packet. But as Cerf and Kahn noted, doing so is prone to errors and risks making the communication non-transparent to the sending host. Instead, the home agent encapsulates these packets in another IP packet addressed to the foreign agent where the mobile host is currently located. That way the application receiving the datagram does not know that the datagram was forwarded by the home agent. Once the foreign agent decapsulates the packet, it cannot simply send it to the address contained in the IP header. That would cause the packets to be routed back to the home network. Instead, it checks to see if the packets are addressed to a mobile host that has registered locally and routes the packets to the mobile host.

Mobile IP thus requires that the network perform three distinct functions:

- A protocol by which mobile nodes can register and deregister with foreign agents.
- A protocol by which foreign agents can notify home agents where the mobile node is currently located.
- Protocols for home agents and foreign agents to encapsulate and decapsulate datagrams they receive.

Unfortunately, this approach suffers from a number of well-known inefficiencies and issues relating to security, handoffs and triangle routing.

1. Security

The ability to register from remote locations raises major security concerns. For example, a malicious user could attempt to mislead the home agent into thinking it was the proper recipient. If so, it could receive all of the packets addressed to the IP address.²⁰¹ Although the architects considered making security a basic feature of IPv6, they eventually decided against doing so.

2. Handoffs

Mobile IP also must find a way to manage the network when a mobile host moves from one base station to another. One solution is to update the home agent. Any tardiness in the update can cause packets to become lost. Another solution is to designate the first foreign agent in a particular transaction as the *anchor foreign agent* that will be the location where the home agent will send all packets. Should the mobile host shift to a different foreign network, the anchor foreign agent can forward the packets to the new location.

3. Triangle Routing

By envisioning that all traffic will travel to the home agent and then be forwarded to the foreign agent, mobile IP employs a form of indirect routing that can be very inefficient. For example, when a person with a home network located in Philadelphia travels to Los Angeles, any packets sent to her while she is in Los Angeles will have to travel across the country to the home agent located in Philadelphia and then be rerouted back to Los Angeles. This can result in the inefficiency of what is sometimes called “triangle routing.”²⁰²

201. See KUROSE & ROSS, *supra* note 107 at 556; PETERSON & DAVIE, *supra* note 104, at 294; TANENBAUM & WEATHERALL, *supra* note 98, at 488.

202. PETERSON & DAVIE, *supra* note 104, at 293.

The home agent can eliminate triangle routing by passing the mobile host's current location on to the sender so that the sender may forward subsequent packets to it directly. The initial communications must still bear the inefficiency of triangle routing. Moreover, such solutions become much more difficult to implement if the mobile agent is constantly on the move.²⁰³ The network must have some way to notify the correspondent that the mobile host has changed location. The usual solution is that much as the home network and the foreign network have agents, the correspondent attempting to contact the mobile host also has a *correspondent agent*. The correspondent agent queries the home agent to learn the location of the mobile host. It then encapsulates the datagram in a new datagram addressed to the foreign agent. The foreign agent then decapsulates the new datagram and passes the original datagram to the mobile host.

The problem arises if the mobile host moves from one foreign network to another. Under indirect routing, the mobile host simply notifies its home agent of the change of location. Under direct routing, however, the correspondent agent is responsible for encapsulating datagrams and forwarding them to the mobile host, not the home agent. At this point, the mobile node needs a way to update the correspondent agent as to its new location. This in turn requires two more protocols:

- A protocol by which correspondent agents can query the home agent as to the mobile node's current location.
- A protocol by which the mobile host that changes foreign networks can notify the correspondent agent about its new location.

The additional complexity is sufficiently difficult to implement that direct routing was not included in the upgrade to IPv6. The net result is that modern mobile broadband networks employ far more intelligence in their core than the end-to-end argument would suggest.

VII. CONCLUSION

The limited ability to add more spectrum and the absolute limit to density of people who can use wireless phones in the same location means that mobile broadband networks must manage their traffic much more aggressively than fixed broadband networks. As noted above, wireless networks often prioritize time-sensitive applications such as voice over non-time-sensitive applications such as email. In addition, certain

203. COMER, *supra* note 104, at 339–46; KUROSE & ROSS, *supra* note 104, at 559–63; TANENBAUM & WEATHERALL, *supra* note 98, at 386–89, 485–88.

solutions, such as the one being advanced by T-Mobile's Binge On, may reduce network congestion, but must do so in an application-specific manner. Bad handoffs, local congestion, and the physics of wave propagation necessitate that mobile broadband networks are subject to highly variable quality of service that requires introducing greater intelligence into the network. The greater heterogeneity of devices and differences in networking standards in the mobile broadband world also limits the feasibility of the prohibition against blocking or throttling devices. Finally, the greater complexity of routing in wireless networks requires introducing a greater degree of intelligence in the core of the network.

The net result is that mobile wireless broadband networks operate on principles that are quite different from those governing the rest of the Internet. Bandwidth limitations require that wireless providers manage their networks more intensively than those operating networks based on other technologies. Because many smartphones do not have IP addresses and wireless networks suffer higher rates of packet loss than fixed networks, wireless broadband networks need to employ virtual circuits and embed intelligence in the network to a greater extent than fixed broadband networks. The unpredictability of signal strength resulting from the physics of wave propagation can necessitate more extensive supervision than other technologies require, as do the realities of system conflicts and power consumption. Lastly, mobility is placing pressure on the routing and addressing space that may soon require more fundamental changes. The industry has not yet reached consensus on the best approach for addressing all of these concerns. In its consideration of regulatory interventions, the FCC must be careful to create a regime that takes these differences into account.

COPYRIGHT REFORM AND COPYRIGHT MARKET: A CROSS-PACIFIC PERSPECTIVE

Jiarui Liu[†]

ABSTRACT

Policymakers around the world are working to unlock copyrighted works from substantial transaction costs in the digital environment. Copyright reform initiatives largely follow one of two directions. Policymakers may change copyright from an opt-in regime into an opt-out or all-in regime (e.g., extended collective licenses and compulsory licenses) to eliminate the necessity of copyright transactions and allow downstream users to exploit copyrighted works without authorization. Alternatively, policymakers may streamline private transactions in the marketplace, create incentives for authors to provide licensing information, and eventually allow market players to innovate on efficient business models. In comparison to the market approach, compulsory licenses have a number of drawbacks (e.g., divesting authors of exclusive rights in copyrighted works, resulting in wasteful rent seeking, and setting arbitrary prices for copyright royalties). However, the fundamental concern is that compulsory licenses would undermine the incentives for collecting societies and other market players to improve their services in order to decrease transaction costs. While the United States and the rest of the world are at a crossroads in copyright reform, the road taken (and the road not taken) by Chinese policymakers provides a valuable lesson: We cannot, in the name of lowering transaction costs, completely sidestep transactions and sidestep the market as the principal mechanism to allocate social resources for intellectual creation.

DOI: <https://dx.doi.org/10.15779/Z38SB3WZ05>

© 2016 Jiarui Liu.

[†] Assistant Professor of Law, University of San Francisco School of Law; Fellow, Center for Internet and Society, Stanford Law School. I am very grateful to Paul Goldstein, Mark A. Lemley, Lawrence M. Friedman, Deborah Hensler, A. Mitch Polinsky, Susan Richey, John Orcutt, Marcus Hurn, Peter K. Yu, Zhou Lin, Glynn S. Lunney, Christopher Jon Sprigman, Ben Depoorter, Kristelia A. García, Guy A. Rub, and all the participants in the Intellectual Property Scholars Conference, the Works-in-Process Intellectual Property Conference, and the Law and Economic Seminar at Stanford Law School for their insightful comments. Of course, all errors remain my own.

TABLE OF CONTENTS

I.	INTRODUCTION: A TALE OF TWO COPYRIGHT BACKLASHES	1462
II.	EXTENDED COLLECTIVE LICENSE	1471
A.	THE TRADITIONAL MODEL	1471
B.	THE CHINESE PROPOSAL	1473
C.	MARKET ALTERNATIVES	1481
III.	COMPULSORY LICENSE	1484
A.	COPYRIGHT AND MONOPOLY	1488
B.	TRANSACTION COSTS AND RENT SEEKING COSTS	1492
C.	POLITICAL COMPROMISE	1496
IV.	ORPHAN WORKS AND MASS DIGITIZATION	1501
A.	LIMITATION ON LIABILITY VERSUS COMPULSORY LICENSE	1503
B.	THE INCENTIVE ANALYSIS	1505
C.	MASS DIGITIZATION	1509
V.	CONCLUSION	1513

I. INTRODUCTION: A TALE OF TWO COPYRIGHT BACKLASHES

In 2012, there were massive public backlashes against copyright reform bills in both the United States and China. However, the two backlashes pointed to dramatically different directions. In the United States, the Stop Online Piracy Act (“SOPA”)¹ and the PROTECT IP Act (“PIPA”)² were proposed in Congress, designed, inter alia, to enlist the assistance of Internet service providers to block overseas websites engaged in copyright infringement. Concerned that the two bills, if passed, could significantly affect Internet governance, leading websites—including Wikipedia, Google, and Twitter—launched an online protest against SOPA and PIPA by temporarily blacking out their services or front pages simultaneously on January 18th. Many Internet users quickly followed suit and sent millions

1. *Stop Online Piracy Act: Hearing on H.R. 3261 Before the H. Comm. on the Judiciary*, 112th Cong. (2011).

2. PREVENTING REAL ONLINE THREATS TO ECONOMIC CREATIVITY AND THEFT OF INTELLECTUAL PROPERTY (PROTECT IP) ACT OF 2011, S. REP. NO. 112-39, at 6–14 (2011).

of messages to their representatives in Congress to express their opposition.³

The concurrent backlash in China, resulting from the 2012 Chinese Copyright Reform Bill, was of a totally different nature.⁴ First, instead of opposing copyright expansion, the public was protesting against copyright limitation, specifically compulsory licenses that would degrade exclusive rights in copyrighted works to rights of remuneration.⁵ Second, Internet companies or users did not initiate the campaign. Instead, Chinese musicians, artists, and authors led it. They expressed their views in newspapers, blogs, and microblogs, filed petitions to the Chinese government, and debated the relevant issues with government representatives on TV shows.⁶ During the one-month period of public inquiries for the new bill, the National Copyright Administration of China (“NCAC”) received over 1,600 formal petitions and found more than one million comments posted on its official website.⁷ Most remarkably, in a country with a 90% piracy rate, the general public appeared to be highly sympathetic to authors and applaud their efforts against rent seeking and

3. For detailed discussions, see Jonathan Band, *The SOPA-TPP Nexus*, 28 AM. U. INT’L L. REV. 31, 42–43 (2012-2013); Yafit Lev-Aretz, *Copyright Lawmaking and the Public Choice: From Legislative Battles to Private Ordering*, 27 HARV. J.L. & TECH. 203, 222 (2013); Yochai Benkler, Hal Roberts, Robert Faris, Alicia Solow-Niederman & Bruce Etling, *Social Mobilization and the Networked Public Sphere: Mapping the SOPA-PIPA Debate*, BERKMAN CENTER FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY (July 25, 2013), http://cyber.law.harvard.edu/publications/2013/social_mobilization_and_the_networked_public_sphere.

4. *Notice of Inquiries Regarding the Draft Amendment to the Copyright Law of the People’s Republic of China*, NATIONAL COPYRIGHT ADMINISTRATION OF CHINA (Mar. 31, 2012), <http://www.ncac.gov.cn/chinacopyright/contents/483/17745.html> [hereinafter “First Draft”]; *Notice of Inquiries Regarding the Second Draft Amendment to the Copyright Law of the People’s Republic of China*, NATIONAL COPYRIGHT ADMINISTRATION OF CHINA (July 6, 2012), <http://www.ncac.gov.cn/chinacopyright/contents/483/17753.html> [hereinafter “Second Draft”]; *Notice of Inquiries Regarding the Draft Amendment for Review to the Copyright Law of the People’s Republic of China*, LEGISLATIVE AFFAIRS OFFICE OF THE STATE COUNCIL (June 6, 2014), <http://www.chinalaw.gov.cn/article/cazjgg/201406/20140600396188.shtml> [hereinafter “Third Draft”].

5. See e.g., Sina Entertainment, *Copyright Amendment Bill Causes Controversies: Gao Xiaosong and Song Ke Opposed on Weibo*, SINA NEWS (Apr. 3, 2012), <http://ent.sina.com.cn/y/n/2012-04-05/12163597772.shtml>; *Record Industry Commission Requests that Section 60 and 70 of the Draft Copyright Law Be Eliminated*, CHINA NEWS (Apr. 28, 2012), <http://www.chinanews.com/cul/2012/04-28/3853299.shtml>.

6. See National Copyright Administration of China (NCAC), *Annual Report of the Third Copyright Reform in China*, G. ADMIN. PRESS & PUBL’N (Nov. 16, 2012), <http://www.gapp.gov.cn/govpublic/96/116997.shtml>.

7. See *id.*

government taking.⁸ The leading traditional and Internet media, controlled primarily by the Chinese government, produced extensive coverage on the controversies surrounding copyright reform, again surprisingly in a rather impartial way.⁹ For those familiar with Chinese media, it is a rare treat to see headlines calling a government initiative “controversial,” “under fire,” and “muzzling.”¹⁰

The NCAC has unprecedentedly prepared three different revised drafts in response to the public outcry over earlier versions.¹¹ Over 90% of the proposed provisions in the first draft were abolished or revised in the subsequent drafts. Among other things, the compulsory license for mechanical reproductions of musical works has been eliminated¹² and the scope of the proposed extended collective license (“ECL”) has been narrowed significantly.¹³

In the meantime, copyright laws in many other countries, including the United States, Canada, and European Union member nations, are also undergoing profound changes in the wake of the rapid advance of digital technologies. Various proposals for digital copyright reform generally revolve around four different forms of legal entitlement: (i) an opt-in regime, with traditional exclusive rights,¹⁴ where an author may prevent the use of her copyrighted works unless she has granted a license; (ii) an opt-out regime where an author may not prevent the use of her copyrighted

8. For detailed discussions of copyright piracy in China, see generally Jiarui Liu, *Copyright for Blockheads: An Empirical Study of Market Incentive and Intrinsic Motivation*, 38 COLUM. J.L. & ARTS 467 (2015); Eric Priest, *Copyright Extremophiles: Do Creative Industries Thrive or Just Survive in China's High-Piracy Environment?* 27 HARV. J.L. & TECH. 467 (2014); Peter K. Yu, *From Pirates to Partners: Protecting Intellectual Property in China in the Twenty-First Century*, 50 AM. U. L. REV. 133 (2000).

9. See, e.g., Xinhua, *Controversial Copyright Amendment to See Legislative Review* (Nov. 01, 2012), <http://en.people.cn/90785/7999907.html>; Lu Na, *New Copyright Law Amendment under Fire* (Apr. 6, 2012), http://www.china.org.cn/arts/2012-04/06/content_25077760.htm; Ember Swift, *Law Revision Could Muzzle China's Music Industry* (Apr. 26, 2012), http://www.china.org.cn/opinion/2012-04/26/content_25247058.htm.

10. See, e.g., Xinhua, *Controversial Copyright Amendment to See Legislative Review*, PEOPLE'S DAILY ONLINE (Nov. 01, 2012), <http://en.people.cn/90785/7999907.html>; Lu Na, *New Copyright Law Amendment under Fire*, CHINA.ORG.CN (Apr. 6, 2012), http://www.china.org.cn/arts/2012-04/06/content_25077760.htm; Ember Swift, *Law Revision Could Muzzle China's Music Industry*, CHINA.ORG.CN (Apr. 26, 2012), http://www.china.org.cn/opinion/2012-04/26/content_25247058.htm.

11. See *supra* note 4 and accompanying text.

12. See *id.*

13. See *id.*

14. See, e.g., WIPO Copyright Treaty, Dec. 20, 1996, S. TREATY DOC. NO. 105-17, 36 I.L.M. 65 (“WCT”); WIPO Performances and Phonograms Treaty, Dec. 20, 1996, S. TREATY DOC. NO. 105-17, 36 I.L.M. 76 (“WPPT”).

works unless she takes affirmative steps to allege her rights, such as ECLs,¹⁵ notice-and-takedown procedures,¹⁶ and traditional formalities;¹⁷ (iii) an all-in regime where an author may not prevent the use of her copyright works, but is entitled to reasonable remuneration in the form of compulsory licenses¹⁸ or public levies;¹⁹ and (iv) an all-out regime where fair use and other exemptions allow third parties to use copyright works without permission and without paying compensation.²⁰

Notably, in North America and the European Union, compulsory licenses and ECLs receive widespread support as potential solutions to the thorny questions regarding orphan works, mass digitization, and online music.²¹ This enthusiasm for regulation and collectivism in digital copyright dramatically contrasts the general attitude of the Chinese public. China, often viewed as a pirate kingdom in the eyes of international observers,²² is apparently moving towards strengthening exclusive rights and removing roadblocks to the formation of efficient copyright markets.

Such parallel developments provide us with a golden opportunity to evaluate how compulsory licenses and ECLs fare for modern creative industries in the digital age. These extra-market mechanisms are often

15. For detailed introductions of the ECL, see World Intellectual Property Organization [WIPO], *Annotated Principles of Protection of Authors, Performers, Producers of Phonograms and Broadcasting Organizations in Connection with Distribution of Programs by Cable*, 20 COPYRIGHT 131, 151 (1984); H. Lund Christiansen, *The Nordic Licensing Systems – Extended Collective Agreement Licensing*, 13 E.I.P.R. 346, 349 (1991); Daniel Gervais, *Application of an Extended Collective Licensing Regime in Canada: Principles and Issues Relating to Implementation* (Dept. of Canadian Heritage, 2003), at 5, http://aix1.uottawa.ca/~dgervais/publications/extended_licensing.pdf.

16. See 17 U.S.C. § 512 (2012).

17. See generally Christopher Sprigman, *Reform(aliz)ing Copyright*, 57 STAN. L. REV. 485 (2004) (suggesting the U.S. reintroduce copyright formalities to decrease transaction costs).

18. 17 U.S.C. §§ 111, 112(e), 114(f), 115(b)(1), 118(b), 119, 122 (2012).

19. 17 U.S.C. § 1003 (2012).

20. See, e.g., *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014).

21. See, e.g., Enterprise and Regulatory Reform Act, 2013, c.24, § 77(3) (U.K.) (amending The Copyright, Designs and Patents Act of 1988) (inserting §§ 116A(1), 116B(1)); U.S. COPYRIGHT OFFICE, ORPHAN WORKS AND MASS DIGITIZATION 6 (June 2015), <http://copyright.gov/orphan/reports/orphan-works2015.pdf> [hereinafter 2015 Orphan Works Study]; Volker Ralf Grassmuck, *A Copyright Exception for Monetizing File-Sharing: A Proposal for Balancing User Freedom and Author Remuneration in the Brazilian Copyright Law Reform* (2010), https://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/Rethinking_Music_Copyright_Exception_Monetizing_File-Sharing.pdf.

22. See generally LOKE-KHOON TAN, *PIRATES IN THE MIDDLE KINGDOM: THE ENSUING TRADEMARK BATTLE* 30 (2007) (explaining why piracy and counterfeiting are rampant in China).

praised for diminishing transaction costs and facilitating large-scale uses of copyright works while providing authors with equitable compensation, which is believed to reflect a middle ground between technology providers and copyright owners.²³ However, empirical evidence suggests that private parties continue to negotiate with each other in the wake of compulsory licenses.²⁴ A widely celebrated example is the Harry Fox Agency, which handles the majority of U.S. mechanical licenses despite the compulsory licenses under 17 U.S.C. § 115.²⁵ In many cases, not only do compulsory licenses fall short of eliminating transaction costs, but they also set in motion wasteful rent seeking to lobby the government for favorable terms.²⁶ The aggregation of transaction costs and rent-seeking costs could render compulsory licenses more burdensome than private transactions in the marketplace.

More importantly, governmental rate-setting procedures are unlikely to produce equitable remuneration for authors in proportion to the market values of their works.²⁷ Digital technology has lowered the market entry barriers of copyright industries and bred an increasing variety of business models for authors. Taking the music industry as an example, some musicians distribute their music in à-la-carte stores like iTunes, some license to subscription-based services like Spotify (a music buffet offering unlimited access to a large repertoire), and others provide music for free on YouTube or even at a negative price (i.e., paying broadcasters to perform their works) to obtain advertising revenues and promote their record or concert sales.²⁸ It is unclear how a uniform price imposed on all relevant authors by the government or another centralized organization could

23. See *infra* note 170 and accompanying text.

24. See generally Mark A. Lemley, *Contracting around Liability Rules*, 100 CALIF. L. REV. 463 (2012) (discussing examples of mechanical license, cable retransmission, and fair use).

25. See 17 U.S.C. § 115(c)(3)(H)(5) (2012); Lydia Pallas Loren, *Untangling the Web of Music Copyrights*, 53 CASE W. RES. L. REV. 673, 682 n.38 (2003) (“The preference for obtaining licenses from Harry Fox instead of utilizing the statutory license is largely due to the reduction of transaction costs offered by Harry Fox.”); HARRY FOX AGENCY, *Licensing General FAQ*, <https://secure.harryfox.com/public/Licensing-GeneralFAQ.jsp> (describing the business scope of the Harry Fox Agency).

26. See *infra* text accompanying note 177.

27. See *infra* note 210 and accompanying text.

28. For discussions of nonmonetary motivations, see, for example, Molly Shaffer Van Houweling, *Making Copyright Work for Authors Who Write to Be Read*, 38 COLUM. J.L. & ARTS 381, 383 (2015); Bernard Lang, *Orphan Works and the Google Book Search Settlement: An International Perspective*, 55 N.Y.L.S. L. REV. 111, 131 (2010); David Throsby, *A Work-Preference Model of Artist Behaviour*, in CULTURAL ECONOMICS AND CULTURAL POLICIES 69 (Alan T. Peacock & Ilde Rizzo eds., 1994).

possibly capture the richness of business models that the copyright market would otherwise cultivate. Transforming exclusive rights to rights of remuneration would override the freedom of authors with different aspirations to exploit their works in different ways. Furthermore, absolving copyright liabilities towards authors with a tariff channeled to a government agency or a collective unrelated to the actual authors breaks the connections between authors and their works and between authors and their audiences.²⁹ The collectivism approach rendering works of authorship faceless may significantly affect the public respect for copyright protection in the long run.

In particular, Nordic countries originally designed ECLs, which allow an agreement between a collecting society and its users to be binding on nonmember authors, as an alternative to compulsory licenses in the narrow areas stipulated under the Berne Convention³⁰ and the TRIPS Agreement.³¹ Theoretically, ECLs may have minimal distorting effect on the copyright market where the collecting society is highly representative and efficient.³² There is an obvious irony however: in reality, the strongest proponents of ECLs are more often than not nascent collecting societies that have yet to accumulate sufficient memberships in a market dominated by a large number of foreign or otherwise nonmember works.³³ Currently, various market solutions to large-scale uses of copyrighted works have evolved in lieu of ECL regimes to diminish transaction costs and promote dynamic competition. In the United States, prospective users can obtain virtually

29. See Paul Goldstein, *Copyright*, 55 LAW & CONTEMP. PROBS. 79, 79-80 (1992) (indicating copyright sustains the very heart and essence of authorship by enabling direct communication between authors and audiences).

30. Berne Convention for the Protection of Literary and Artistic Works, Sept. 9, 1886, 1161 U.N.T.S. 30, as amended on Sept. 28, 1979, S. TREATY DOC. NO. 99-27 [hereinafter Berne Convention].

31. Agreement on Trade-Related Aspects of Intellectual Property Rights, Apr. 15, 1994, Marrakesh Agreement Establishing the World Trade Organization, Annex 1C, 1869 U.N.T.S. 299, 33 I.L.M. 1197 [hereinafter TRIPS Agreement].

32. See Stanley M. Besen, Sheila N. Kirby & Steven C. Salop, *An Economic Analysis of Copyright Collectives*, 78 VA. L. REV. 383, 383 (1992) (praising the fact that “copyright collectives lower collection costs and permit more transactions to occur”).

33. See Tarja Koskinen-Olsson, *Collective Management in Nordic Countries*, in COLLECTIVE MANAGEMENT OF COPYRIGHT AND RELATED RIGHTS 283, 306 (Daniel J. Gervais ed., 2010) (“[The system of ECLs in Nordic countries] presupposes in other words that the ‘copyright market’ is well organized and disciplined.”); Johan Axhamn & Lucie Guibault, *Cross-Border Extended Collective Licensing: A Solution to Online Dissemination of Europe’s Cultural Heritage?* (Instituut Voor Informatierecht, 2011), at viii, http://www.ivir.nl/publicaties/guibault/ECL_Europeana_final_report092011.pdf (“[The ECL] presupposes the existence of a representative CMO with a sound culture of good governance and transparency.”).

total freedom to use any musical work by purchasing blanket licenses in a competitive market of collecting societies that include ASCAP, BMI, and SESAC.³⁴ In China, certain collecting societies offer an indemnification clause in their licenses, defending and holding their users harmless from copyright claims by nonmembers.³⁵ Hence, the fundamental concern about ECLs is that collecting societies who strive to increase membership may be incentivized to lobby the government for more powers rather than to attract more members by providing better services and increasing operational efficiency.

A compulsory license has also been introduced in several countries as a legislative solution to the problem of orphan works for which copyright owners may not be identified or located with a diligent search.³⁶ A user may apply to a government agency for the compulsory license to use an orphan work upon payment to a collecting society designed by the government agency. However, the “compulsory license” approach is clearly less efficient than the “limitation on liability” approach proposed by the U.S. Copyright Office, which would allow users to use orphan works without payment until copyright owners emerge.³⁷ First, the compulsory license is rarely used in the countries where it has been enacted³⁸ because it would require substantial administrative costs to certify the legal status of orphan works case-by-case, determine the royalty rates for the compulsory license, and process upfront payments by users to a collecting society. Second, it proves difficult for a government agency to efficiently set royalty rates for orphan works, which naturally do not have current market benchmarks.³⁹ Third, the compulsory license may create an incentive for the collecting society to sit on royalty payments rather than to diligently search for copyright owners if it permits the collecting society to retain the unallocated amount to defray administrative costs and support collective-purpose projects for existing members.⁴⁰ Fundamentally, copyright protection

34. See U.S. COPYRIGHT OFFICE, COPYRIGHT AND THE MUSIC MARKETPLACE 19 (Feb. 2015), <http://copyright.gov/policy/musiclicensingstudy/copyright-and-the-music-marketplace.pdf> [hereinafter Music Marketplace Study].

35. Interviews with QJM & YDK, executives in Chinese collecting societies (Nov. 2, 2010) (on file with the author).

36. See, e.g., Copyright Act, R.S.C. 1985, c. C-42, s. 77 (Can.).

37. See U.S. COPYRIGHT OFFICE, REPORT ON ORPHAN WORKS (Jan. 2006), <http://copyright.gov/orphan/orphan-report-full.pdf> [hereinafter 2006 Orphan Works Study].

38. See *infra* note 248 and accompanying text.

39. See *infra* note 251 and accompanying text.

40. See *infra* note 255 and accompanying text.

reflects a trade-off between incentive and access.⁴¹ However, orphan works by definition involve authors who may not be located with a diligent search. So chances are the authors will never reappear. If compulsory licenses are imposed in these cases, users would pay a higher price, but the real authors would not receive any financial incentives. This situation would be the worst of both worlds: limited access for consumers *and* no incentives for authors.

Notably, the clearance difficulties concerning mass digitization projects do not result entirely from orphan works. For instance, only a quarter of the Google Books Project corpus are potentially orphan works.⁴² Neither does the sheer volume of copyrighted works involved in the Google Books Project by itself justify a statutory exemption. The increase in transaction costs has been approximately proportionate to the increased volume and increased value of the overall database. It does not make much sense to categorically argue that the more copyrighted works a database contains, the less reasonable it is to require a copyright license.

The key barrier to mass digitization appears to be that the incremental value of any individual work to the whole project is often lower than the transaction cost needed to obtain a license for such a work. However little it takes to locate a copyright owner (say one dollar), the user would still not reach out for a license if scanning the book adds even less (say three cents) to the project.⁴³ This issue is similar to the clearance hurdle for television/radio broadcasters who perform a significant number of musical works for their daily programs. If history is any indication, the best solution is not to bypass copyright transactions. Instead, we may pool different copyrighted works together through major publishers or collecting societies to facilitate the issuance of blanket licenses for mass digitization. Meanwhile, the ECL model does not appear to be necessary here. A project like Google Books does not need to include all books in the world to become a viable business, as the holdup problem rarely arises in the context of mass

41. For detailed surveys of economic theories in connection with copyright law, see PAUL GOLDSTEIN, *GOLDSTEIN ON COPYRIGHT* §1 (2015); RICHARD POSNER & WILLIAM LANDES, *THE ECONOMIC STRUCTURE OF INTELLECTUAL PROPERTY LAW* 37 (2003); Gillian K. Hadfield, *The Economics of Copyright: An Historical Perspective*, 38 *COPYRIGHT L. SYMP. (ASCAP)* 1 (1992).

42. See, e.g., Michael Cairns, *580,388 Orphan Works—Give or Take*, PERSONANONDATA (Sept. 9, 2009), <http://personanondata.blogspot.com/2009/09/580388-orphan-works-give-or-take.html>.

43. See *infra* note 269 and accompanying text.

digitization.⁴⁴ If Google has obtained blanket licenses from collecting societies but a nonmember author refuses to grant her individual license, it may simply delete the infringing work from the database and continue the operation with other licensed works. A single party can scarcely have any veto power to block the entire project. As a matter of fact, Google has indicated that it only scanned one fourth of the books in the world even though a federal court has legalized the project as a fair use for book searching and non-display purposes.⁴⁵

This Article does not aim to simply document ongoing copyright reform discourses in the United States and China, but rather to assess the rationality of compulsory licenses and ECLs vis-à-vis market mechanisms in modern copyright industries. Nevertheless, the discussions indeed shed light on the cultural contexts useful to understand why the Chinese public appears to be more conservative about compulsory licenses. Compulsory licenses and similar mechanisms are often proposed to decrease transaction costs. However, they also run the risk of depriving authors of pricing rights and replacing the copyright market with a small group of centralized decision makers. If lowering transaction costs is the most important goal for a copyright regime, the Chinese public has actually experienced first-hand a world without any transaction costs: all producers sold their products to the government and all consumers bought their products from the government. No one had the right to determine prices and terms except for a small group of government planners. This was what we call the central-planning economy. The Chinese people tried it once and didn't like it. Thus, they moved on to the market economy, which gave rise to the economic growth of China in the last three decades.⁴⁶ This experience teaches all of us a valuable lesson: we cannot, in the name of lowering transaction costs,

44. See, e.g., Guido Calabresi & A. Douglas Melamed, *Property Rules, Liability Rules, and Inalienability: One View of the Cathedral*, 85 HARV. L. REV. 1089, 1107 (1972). Similar issues are sometimes called the tragedy of "anticommons." See Michael A. Heller, *The Tragedy of the Anticommons: Property in the Transition from Marx to Markets*, 111 HARV. L. REV. 621, 623 (1998); Michael A. Heller & Rebecca S. Eisenberg, *Can Patents Deter Innovation? The Anticommons in Biomedical Research*, 280 SCIENCE 698, 700 (1998).

45. Leonid Taycher, *Inside Google Books: Books of the World, Stand Up and Be Counted! All 129,864,880 of You.*, GOOGLE: INSIDE SEARCH BLOG (Aug. 5, 2010, 8:26 AM), <http://booksearch.blogspot.com/2010/08/books-of-world-stand-up-and-be-counted.html>; Jennifer Howard, *Google Begins to Scale Back Its Scanning of Books from University Libraries*, THE CHRONICLE OF HIGHER ED. (Mar. 9, 2012), <http://chronicle.com/article/Google-Begins-to-Scale-Back/131109/>.

46. See *World Economic Outlook Database*, INT'L MONETARY FUND (July 2015), <https://www.imf.org/external/pubs/ft/weo/2015/01/weodata/index.aspx>.

completely sidestep transactions and sidestep the market as the principal mechanism to allocate social resources for intellectual creation.

Part II begins with an introduction of the ECL system in Nordic countries. It demonstrates that the Chinese proposal would likely transgress the traditional contours of the Nordic model, undermine incentives for collecting societies to improve their operational efficiency, and violate the three-step test under the Berne Convention and the TRIPS Agreement. Part III evaluates the rationality of compulsory licenses in the digital age. It contends that compulsory licensing is neither necessary to prevent monopoly nor effective in lowering transaction costs, and that it often results from a political compromise that unduly deflates the market values of copyrighted works. Part IV compares the “limitation on liability” and “compulsory license” approaches as possible solutions to orphan work and mass digitization issues. It explains why the former is economically superior. Part V concludes the article with a summary of the major points.

II. EXTENDED COLLECTIVE LICENSE

A. THE TRADITIONAL MODEL

The traditional extended collective license (“ECL”), as pioneered in Nordic countries since the 1960s, normally includes four features.⁴⁷ First, the ECL allows the agreement concluded by a collecting society and users to extend to and become binding on nonmembers by operation of law.⁴⁸ In other words, once the licensing agreement officially takes effect, the licensee would have the legal privilege to use the entirety of copyright works in the category that the collecting society administers, including those owned by nonmember authors. Members and non-members would have equal rights to request remuneration from the collecting society under the licensing agreement.

Second, a collecting society with the power to issue the ECL must obtain authorization from the government and represent a substantial

47. See *supra* note 15 and accompanying text.

48. See Annette Dilley & Thomas Dyekjær, *The Danish Copyright System and Copying Within Libraries*, 31 INT’L J. LEGAL INFO. 445, 449 (2003); Thomas Riis & Jens Schovsbo, *Extended Collective Licenses and the Nordic Experience – It’s a Hybrid but Is It a Volvo or a Lemon?*, 33 COLUM. J.L. & ARTS 471, 496 (2010); Henry Olsson, *The Extended Collective License as Applied in the Nordic Countries*, KOPINOR (Mar. 10 2010), <http://www.kopinor.no/en/copyright/extended-collective-license/documents/the-extended-collective-license-as-applied-in-the-nordic-countries>.

number of right holders of the category involved.⁴⁹ Nordic countries usually do not provide for a clear definition or threshold for “a substantial number of right holders.” But it is generally understood that, although it may be difficult numerically to verify whether a collecting society represents the majority of right holders of a category, close to half is required.⁵⁰

Third, the ECL is typically limited to special areas of usage, including: (i) retransmission by cable or satellite of television and radio programs;⁵¹ (ii) use of published works by certain television and radio stations;⁵² (iii) reproduction for research, educational and internal purposes;⁵³ (iv) preservation by libraries, archives, and museums;⁵⁴ (v) reproduction for the visually handicapped and the hearing impaired;⁵⁵ and (vi) use of works of fine art displayed in public.⁵⁶ It appears that the ECL principally applies to the areas where, due to exorbitant transaction costs, copyright owners may not be able to individually grant licenses to end users in an effective way.⁵⁷ Therefore, it is unsurprising that substantial overlap exists between the coverage of ECLs and coverage of compulsory licenses permitted in international conventions and implemented in domestic laws.⁵⁸

49. See Thomas Riis & Jens Schovsbo, *Extended Collective Licenses in Action*, 43 I.I.C. 930, 935 (2012).

50. See *id.* at 937.

51. Consolidated Act on Copyright § 35 (2010) (Den.); Copyright Act 1961 (as amended) § 25(f)(4) (Finland); Copyright Act 1961 (as amended) § 34 (Nor.).

52. Consolidated Act on Copyright §§ 30 & 30(a) (2010) (Den.); Copyright Act 1961 (as amended) §§ 25(f),(g) (Fin.); Copyright Act (as amended) §§ 30 & 32 (1961) (Nor.).

53. Consolidated Act on Copyright §§ 13 & 14, (2010) (Den.); Copyright Act §§ 13(a) & 14 (1961) (as amended) (Finland); Copyright Act (as amended) §§ 13(b) & 14 (1961) (Nor.).

54. Consolidated Act on Copyright § 16(b) (2010) (Den.); Copyright Act § 16(d) (1961) (as amended) (Fin.); Copyright Act (as amended) § 16(a) (1961) (Nor.).

55. Consolidated Act on Copyright § 17(4) (2010) (Den.); Copyright Act (as amended) § 17(b) (1961) (Nor.).

56. Consolidated Act on Copyright § 24(a) (2010) (Den.); Copyright Act § 25(a)(2) (1961) (as amended) (Fin.).

57. See Jane Ginsburg, *Reproduction of Protected Works for University Research or Teaching*, 39 J. COPYRIGHT SOC'Y U.S.A. 181, 198 (1992) (“[I]t is difficult to argue that ECL-statutes have deprived authors of a right that, practically speaking, was impossible to enforce prior to the existence of the ECL-statutes.”).

58. See Berne Convention, *supra* note 30, Article 11 bis(2) (rebroadcasting), 9(2) (reproduction); see also Christiansen, *supra* note 15, at 349 (arguing the ECL “is regarded as a compulsory license in the sense of the Berne Convention”).

Finally, nonmember right holders sometimes (but not always) may opt out of the ECL agreement by serving a notice to either the collecting society or the user.⁵⁹

B. THE CHINESE PROPOSAL

The first draft of the 2012 Chinese Copyright Reform Bill included a proposal for the ECL system:

Collective management organizations that have obtained authorization from right holders and are capable of representing right holders' interests nationwide may apply to the Copyright Administration of the State Council for representing the entirety of right holders to exercise their copyrights and related rights, except those right holders who have declined collective management in a written declaration.⁶⁰

While the ECL system has worked fairly well in Nordic countries for over sixty years without much controversy, the Chinese ECL proposal has been confronted with strong opposition. A number of high-profile Chinese musicians and publishers announced that, should the ECL system stay in the final legislation, they would withdraw from the Music Copyright Society of China—the collecting society that was one of the driving forces behind the ECL proposal.⁶¹ They also threatened to establish a competing collecting society of their own.⁶² The mounting pressure from musicians, publishers, and the general public resulted in a dramatic turn of events: The society was forced to issue a public statement that backpedaled its support for the ECL system and called its initial enthusiasm for the ECL proposal

59. See Daniel J. Gervais & Alana Maurushat, *Fragmented Copyright, Fragmented Management: Proposals to Defrag Copyright Management*, 2 CAN. J. L. & TECH. 15, 31 (2003); Stef Van Gompel, *Unlocking the Potential of Pre-Existing Content: How to Address the Issue of Orphan Works in Europe?*, 38 I.I.C. 669, 688 (2007); U.K. INTELLECTUAL PROPERTY OFFICE, © THE WAY AHEAD: A STRATEGY FOR COPYRIGHT IN THE DIGITAL AGE 38 (2009), <http://webarchive.nationalarchives.gov.uk/20140603093549/http://www.ipo.gov.uk/c-strategy-digitalage.pdf>.

60. Article 60, First Draft, *supra* note 4.

61. See Xu Haiyang & Shen Can, *Musicians Collectively Published Their Comments on the Draft Copyright Law*, SINA (Apr. 13, 2012), <http://ent.sina.com.cn/c/2012-04-13/09473604126.shtml>.

62. Apparently, several influential musicians went ahead and established a competing society in 2015, called Hua Yue Music. See HUA YUE MUSIC (BEIJING) CO., www.huayuemusic.com (last visited Apr. 2, 2017).

“too narrow-minded, short-sighted, geeky and insensitive to authors’ interests.”⁶³

The controversy surrounding the ECL proposal is unsurprising given that the model suggested in the first draft of the 2012 Chinese Copyright Reform Bill is substantially different from the Nordic model. First, the proposed ECL is not confined to any special areas. Any collecting society may apply to the NCAC for the legal power to impose the ECL within its scope of business. Currently, there are five collecting societies operating in mainland China: (i) the Music Copyright Society of China (“MCSC”), established in 1992 to manage musical works;⁶⁴ (ii) China Audio-Video Copyright Association (“CAVCA”), established in 2008 to manage music videos used by karaoke bars;⁶⁵ (iii) China Written Works Copyright Society (“CWWCS”) established in 2008 to manage literary works;⁶⁶ (iv) the Images Copyright Society of China (“ICSC”), established in 2008 to manage photographic works;⁶⁷ and (v) China Film Copyright Association (“CFCA”), established in 2009 to manage motion pictures.⁶⁸ Their scopes of business in the aggregate cover almost all commercially significant uses for copyrighted works (except software), much broader than what the Nordic ECL encompasses. Accordingly, the potential impacts of the Chinese proposal on exclusive rights would be much more extensive.

Second, although the proposed ECL contains an opt-out option for nonmembers, the option is essentially superfluous. The first draft of the 2012 Chinese Copyright Reform Bill simultaneously introduces a limitation on liability that would totally negate any financial incentives for nonmembers to opt out of the ECL: a nonmember author would merely be entitled to the royalties collectable from a copyright society if the infringing

63. See various comments posted online by the MCSC on the Draft Amendment: <http://www.mcsc.com.cn/imS-13-994.html> (Apr. 6, 2012); <http://www.mcsc.com.cn/imS-13-998.html> (Apr. 13, 2012); <http://www.mcsc.com.cn/imS-13-1003.html> (Apr. 19, 2012).

64. See MCSC, *Basic Information*, <http://www.mcsc.com.cn/mIL-5.html> (last visited Apr. 2, 2017).

65. See *Association History*, CHINESE AUDIO-VIDEO COPYRIGHT ASSOCIATION, <http://www.cavca.org/xhlc.php?page=28> (last visited Apr. 2, 2017).

66. See *Association Introduction*, CHINA WRITTEN WORKS COPYRIGHT SOCIETY, <http://www.prcopyright.org.cn/staticnews/2010-01-28/100128145635437/1.html> (last visited Apr. 2, 2017).

67. See *Association Introduction*, IMAGES COPYRIGHT SOCIETY OF CHINA, <http://www.icsc1839.org/html/zhuzuoquanxiehui/guanyuxiehui/index.html> (last visited Apr. 2, 2017).

68. See CHINA FILM COPYRIGHT ASSOCIATION (COLLECTIVE), <http://www.cfca-c.org/> (last visited Apr. 2, 2017).

user has obtained a license from the collecting society.⁶⁹ Most authors who prefer to opt out of the ECL presumably believe that they are capable of collecting more royalties by managing their works on their own, probably because they are more efficient in negotiating copyright licenses or they create more value for prospective users. However, the limitation on liability provision would impose a hard cap on how much an author could possibly collect. Even if she has opted out of the ECL, she could demand no more than what she would otherwise receive from the collecting society. In that case, who would bother to opt out?⁷⁰

Third, the most important difference between the Chinese proposal and the Nordic ECL is not immediately obvious: all the existing Chinese collecting societies are far from the level of representativity envisioned by the Nordic tradition. As Figure 1 shows below, the MCSC, the largest and oldest collecting society in China, has only 6,500 members. It amounts to around 10% in scale of ASCAP and BMI in the United States and GEMA in Germany; and it appears even smaller compared to STIM of Sweden and PRS-music in the United Kingdom. This low level of representativity is highly disproportionate to the Chinese population,⁷¹ which is more than a hundred times larger than the Swedish population.

The lack of representativity significantly aggravates the issues inherent in an ECL regime. The ECL may dilute exclusive rights granted by copyright law and undermine creative incentives for nonmember authors.⁷² Several commentators observe that the ECL is not really different from a compulsory license in effect.⁷³ For instance, in Nordic countries, the ECLs

69. Article 70, First Draft, *supra* note 4.

70. Notably, this provision is drastically different from the provision in some countries where an author would be entitled only to what would be available from a collecting society if she is a *member* of the collecting society. See Daniel J. Gervais, *The Purpose of Copyright Law in Canada*, 2 U. OTTAWA L. & TECH. J. 315, 351 (2005).

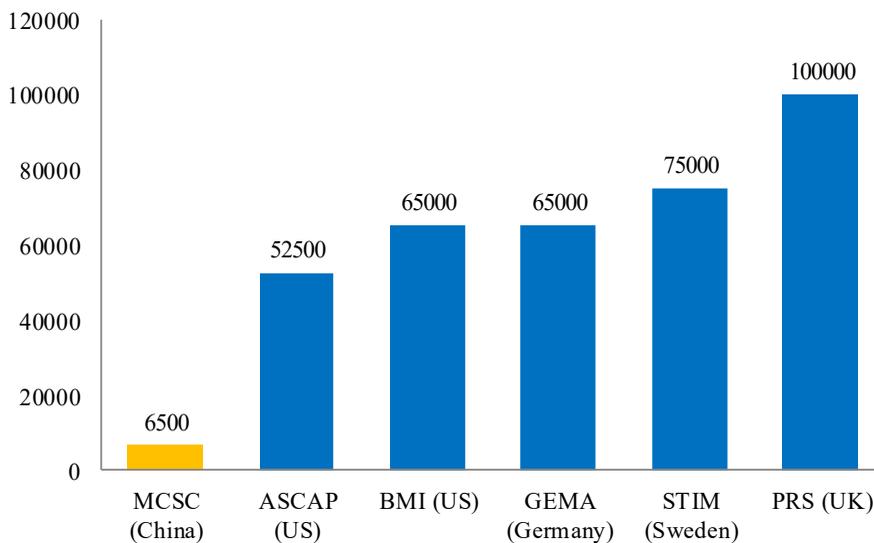
71. By comparison, China Musicians Association, which consists exclusively of established musicians recognized by the government, has 13,900 members. See *Info, CHINESE MUSICIANS ASSOCIATION*, <http://www.chnmusic.org/CmaInfo.html>.

72. See, e.g., Annette Kur & Jens Schovsbo, *Expropriation or Fair Game for All? The Gradual Dismantling of the IP Exclusivity Paradigm* (Max Planck Institute for Intellectual Property, Competition & Tax Law Research Paper No. 09-14, 2009), <http://ssrn.com/abstract=1508330>, at 13.

73. See, e.g., Maria Pallante, *Orphan Works, Extended Collective Licensing and Other Current Issues*, 34 COLUM. J.L. & ARTS 23, 31 (2010) (“Collective extending licensing is basically a form of statutory licensing.”); Riis & Schovsbo, *supra* note 48, at 476.

in areas such as cable transmission are entirely compulsory without an opt-out option for authors.⁷⁴

Figure 1: Memberships of Various Collecting Societies⁷⁵



Furthermore, nonmember authors, especially foreign authors, may not always be aware that their works are used under an ECL agreement. If an author does not join the domestic society or join a foreign society that has a reciprocal agreement with the domestic society, she is practically denied the right to opt out or to claim remuneration. The reason is that a collecting society typically does not have a legal obligation to seek out nonmembers whose works are used.⁷⁶ Neither does it have internal incentives to do so. If the royalties collected remain unclaimed for a certain amount of time (say three years), a collecting society would normally distribute the funds among

74. Council Directive of 27 September 1993 on the Coordination of Certain Rules Concerning Copyright and Rights Related to Copyright Applicable to Satellite Broadcasting and Cable Retransmission, Art. 9(1) 93/83/EEC, 1993 O.J. (L 248) 15 [hereinafter the E.U. Satellite and Cable Directive].

75. Membership data for Figure 1 was obtained from collecting society websites. See <http://www.mcsc.com.cn/imS-13-994.html> (MCSC); *About Us*, ASCAP, <http://www.ascap.com/about/> (last visited April 4, 2017); *About*, BMI, <http://www.bmi.com/about> (last visited April 4, 2017); *About Us*, GEMA, <https://www.gema.de/en/> (last visited April 4, 2017); STIM, <https://www.stim.se/en> (last visited April 4, 2017); *Our Members*, PRS FOR MUSIC, <https://www.prsformusic.com/aboutus/ourorganisation/ourmembers/Pages/default.aspx> (last visited April 4, 2017).

76. See Riis & Schovsbo, *supra* note 48, at 483.

existing members or use them to cover administrative costs.⁷⁷ This windfall suggests that the ECL may financially benefit those collecting societies that sit on the royalties collected for unidentified authors. The prejudice against foreign authors appears more severe when a collecting society decides to withhold certain funds (e.g., 10% of the royalties collected) for social, cultural, and other collective purposes in the forms of stipends and scholarships.⁷⁸ Although the royalties are derived from both domestic and foreign works, typically only domestic authors are provided access to such stipends, a result akin to economic protectionism and thus in violation of the spirit (if not mandates) of national treatment under the Berne Convention and the TRIPS Agreement.⁷⁹

The ECL, by extending the licensed repertoire to all works in a category, also strengthens the monopolistic positions of existing collecting societies against both users and authors. Anecdotes illustrate how monopolistic societies have wielded their market power to squeeze exorbitant royalties from their users. For example, in 1989 the French collecting society SACEM was sued for unfair trading and price fixing because it charged discotheques at a royalty rate fifteen times higher than sister societies, including GEMA, normally charged.⁸⁰ The Canadian collecting society Access Copyright faced vociferous opposition from universities in 2012 when it increased the annual license fee per student to \$26, significantly higher than the \$3.56 per student fee that its sister society Copyright Clearance Center (“CCC”) charged in the United States.⁸¹ In 1992, MTV Europe filed a complaint alleging price fixing by VPL, a collecting society established by IFPI to manage music videos for major music labels in Europe. The labels, through VPL, jointly demanded that MTV Europe pay

77. For methodologies to distribute unclaimed funds, see, for example, *General FAQ*, SOUNDEXCHANGE, <http://www.soundexchange.com/generalfaq/> (last visited Apr. 4, 2017).

78. See Ginsburg, *supra* note 57, at 195 (“The use of collected money for collective purposes is not unique to the Nordic [countries] . . . CISAC have a long tradition of deducting up to 10% of their revenue for various collective purposes.”); Riis & Schovsbo, *supra* note 48, at 492 (“[T]he practice is widespread and generally accepted due to the prevalence of the practice (at least in continental Europe.)”); Ferdinand Melichar, *Deductions Made by Collecting Societies for Social and Cultural Purposes in the Light of International Copyright Law*, 22 I.I.C. 47, 49 (1991).

79. PAUL GOLDSTEIN, *COPYRIGHT’S HIGHWAY: FROM GUTENBERG TO THE CELESTIAL JUKEBOX* 159 (2003).

80. See *Joined Cases 110/88, 241/88 & 242/88, Lucazeau v. SACEM*, 1989 E.C.R. 2811.

81. See Howard Knopf, *AUCC Settlement with Access Copyright—Questions and Answers—or Still More Questions?*, EXCESS COPYRIGHT (May 2, 2012, 7:34 PM), <http://excesscopyright.blogspot.ca/2012/05/aucc-settlement-with-access-copyright.html>.

15% of its gross revenue, while they offered music videos essentially for free in the United States.⁸² The zero-price offered in the United States could indeed reflect the market rate because music companies in a competitive environment would be more open to sacrificing performance royalties, taking into account the obvious effect of television broadcasting to promote record sales.

In theory, the royalty rate without the coordination by a collecting society may even settle at a negative level, as evidenced by the long-term practice of payola, a reverse payment from individual music labels (sometimes music publishers) to radio/television broadcasters in exchange for more airtime of their own works.⁸³ Collecting societies typically grant blanket licenses that charge users uniform prices based on business scales of sales revenue or square footage, regardless of which and how many songs are actually performed. This business model functions in a way similar to a price-fixing cartel that limits price competition between songs and music companies.⁸⁴ However, game theory suggests that, as could happen to any price collusion among oligopolists, individual music companies have an inherent incentive to “cheat,” in other words, to *compete secretly* by offering broadcasters payola under the table to boost music sales and obtain a bigger share of the performance royalty pie. In this sense, payola may be regarded as one piece of evidence that the rates offered by collecting societies are too high and that persistent lobbying efforts by the music industry to outlaw payola actually aim to reinforce price collusion and discipline violators.⁸⁵

Several countries have imposed price control mechanisms, such as governmental agencies in charge of rate-setting proceedings, on collecting societies to protect users from potential supra-competitive pricing. Typical examples include the Copyright Licensing Tribunal in Denmark,⁸⁶ the Copyright Tribunal in the United Kingdom,⁸⁷ and the rate courts at the

82. See Hamish Porter, *European Union Competition Policy: Should the Role of Collecting Societies Be Legitimised?* 18 E.I.P.R. 672, 676 (1996).

83. See KERRY SEGRAVE, *PAYOLA IN THE MUSIC INDUSTRY: A HISTORY, 1880–1991* 36 (1993) (indicating payola payments to radio stations sometimes exceeded the royalties collected by ASCAP).

84. See RICHARD A. POSNER, *ANTITRUST LAW* 30 (2d ed. 2001) (collecting societies are examples of the “benign cartel”); William M. Landes, *Harm to Competition: Cartels, Mergers and Joint Ventures*, 52 *ANTITRUST L. J.* 625, 632 (1983) (calling collecting societies “one of the hallmarks of a successful cartel”).

85. See Ronald H. Coase, *Payola in Radio and Television Broadcasting*, 22 *J.L. & ECON.* 269, 315 (1979).

86. See Riis & Schovsbo, *supra* note 48, at 476.

87. See *About Us*, COPYRIGHT TRIBUNAL, <https://www.gov.uk/government/organizations/copyright-tribunal/about>.

United States District Court for the Southern District of New York.⁸⁸ Governmental rate-setting is notorious for giving rise to wasteful rent seeking, being susceptible to regulatory capture, and systematically underestimating the market values of intellectual products. These drawbacks will be further discussed in the context of various compulsory licenses, but it suffices to point out here, that the widespread imposition of governmental rate-setting in Nordic countries stands in dramatic contrast to the conventional wisdom that ECL regimes can achieve the best of both worlds—decreasing transaction costs *and* safeguarding market transactions.⁸⁹

While the ECL may increase the market power of collecting societies, ample evidence exists indicating that collecting societies with monopolistic positions tend to have less incentive to improve their productive efficiency than those faced with fierce competition in the marketplace. Competitive American collecting societies ASCAP and BMI respectively spent 11.3%⁹⁰ and 11.7%⁹¹ of their royalty revenues to offset administrative costs. By contrast, European collecting societies, most of which traditionally retained *de jure* monopoly within their respective territories, had pocketed 30% to 40% of royalties as administrative fees until market pressure intensified in the mid-1990s and resulted in a sizable decrease of overhead to 15%.⁹² The European Union has recently stipulated a directive on collective rights management for the very purpose of further rejuvenating Europe-wide competition between collecting societies from different member countries.⁹³ Similarly, all of the Chinese collecting societies have monopolistic positions in their respective areas by operation of law.⁹⁴ Their operating

88. Currently, Judge Denise Cote oversees rate-setting proceedings regarding ASCAP, and Judge Louis L. Stanton oversees rate-setting proceedings regarding BMI. *See* Music Marketplace Study, *supra* note 34, at 41.

89. *See* Riis & Schovsbo, *supra* note 48, at 473 (“ECLs have the effectiveness of compulsory licenses but at the same time leave right holders in control of the use of their works.”).

90. *See* ASCAP, <http://www.ascap.com/licensing/licensingfaq.aspx#general>.

91. *See* BMI Tops \$900 Million Mark in Revenues, BMI NEWS (Aug. 25, 2008), http://www.bmi.com/news/entry/bmi_tops_900_million_mark_in_revenues.

92. *See* Jeff Clark-Meads, *U2 Settles Royalty Suit with U.K.’s PRS*, BILLBOARD, Apr. 18, 1998, at 4.

93. Directive 2014/26, of the European Parliament and of the Council of 26 February 2014 on Collective Management of Copyright and Related Rights and Multi-Territorial Licensing of Rights in Musical Works for Online Use in the Internet Market, 2014 O.J. (L 84/72).

94. *See* Article 7, The Regulations of Copyright Collective Management (effective on Mar. 1, 2005), http://www.gov.cn/gongbao/content/2011/content_1860740.htm (allowing no overlapping between the business scopes of different collecting societies).

efficiency is frequently called into question. For instance, collecting societies are semi-governmental organizations in China and their employees actually receive substantial stipends from the Chinese government.⁹⁵ Despite that, CAVCA, the collecting society that manages music videos used by karaoke parlors, was reported to distribute only 46% of the royalties collected to copyright owners (including all composers, performers, and producers), retaining the rest as administrative costs.⁹⁶

In addition, the proposed ECL in China likely runs a higher risk than the Nordic model of violating the three-step test under the TRIPS Agreement: “Members shall confine limitations or exceptions to exclusive rights to certain special cases which do not conflict with a normal exploitation of the work and do not unreasonably prejudice the legitimate interests of the right holder.”⁹⁷ As a WTO panel sensibly explained, the first step of the three-step test, “certain special cases,” requires that limitations or exceptions be well-defined and limited in scope and reach.⁹⁸ The WTO panel went on to find that the business exemption under the U.S. Copyright Act⁹⁹ clearly violates the first step by *not* being limited to “certain special cases” because it effectively exempts 70% of all eating and drinking establishments and 45% of all retail establishments in the United States.¹⁰⁰ In the same vein, if a Chinese collecting society like MCSC has recruited less than 10% of Chinese musicians as its members,¹⁰¹ extending their licenses to the entirety of Chinese musicians would prejudice the exclusive rights of the unassociated 90%, a vast majority that can hardly be reconciled with the requirement of “certain special cases.”

95. See Ye Jiang, *Changing Tides of Collective Licensing in China*, 21 MICH. ST. INT'L L. REV. 729, 744 (2013).

96. See Yifei Tan, *Zhongwenfa Profits from the Karaoke Supervision Platform – Discovering the Distribution Scheme of KTV Revenues*, INFZM.COM (Mar. 25, 2010), <http://www.infzm.com/content/42972/0>.

97. TRIPS Agreement, *supra* note 31, Article 13. The ECL is likely to constitute “exceptions or limitations” under the TRIPS Agreement and the Berne Convention. See *id.*; Berne Convention, *supra* note 30, Article 9(2); Alain Strowel, *Symposium: Collective Management of Copyrights: Solution or Sacrifice? The European “Extended Collective Licensing” Model*, 34 COLUM. J.L. & ARTS 665, 669 (2011) (stating an ECL works in practice as an exception); cf. Séverine Dusollier & Caroline Colin, *Symposium: Collective Management of Copyrights: Solution or Sacrifice? Peer-to-Peer File Sharing and Copyright: What Could Be the Role of Collective Management?*, 34 COLUM. J.L. & ARTS 809, 826 (2011).

98. See Report of the Panel, *United States—Section 110(5) of the U.S. Copyright Act*, ¶ 6.31, WTO Doc. WT/DS160/R (June 15, 2000).

99. See 17 U.S.C. § 110(5) (2012).

100. See Report of the Panel, *supra* note 98, ¶ 6.237.

101. See *supra* note 71 and accompanying text.

C. MARKET ALTERNATIVES

Admittedly, the proposed ECL offers advantages. First, the ECL, similar to any other form of collective management, is designed to minimize transaction costs.¹⁰² By pooling a large number of copyrighted works, collecting societies may theoretically generate economies of scale for authors in negotiating, auditing, and enforcing copyright licenses.¹⁰³

Second, the ECL system would establish a one-stop shop that particularly benefits prospective users who desire to clear all necessary licenses for a large repertoire of copyrighted works in a cost-effective way.

Third, the ECL could immediately and substantially strengthen the bargaining power of collecting societies against large market players.¹⁰⁴ A high-level official at the NCAC once told a story that vividly illustrates this point: In 2010, the Chinese collecting society CWWCS discovered that Google had scanned 210,000 books written by 80,000 Chinese authors. It approached Google and claimed, “We are the only collecting society that manages literary works in China, and you need a license from us to scan

102. See Besen et al., *supra* note 32, at 383.

103. See *Broadcast Music, Inc. v. Columbia Broad. Sys., Inc.*, 441 U.S. 1, 20 (1979): ASCAP and the blanket license developed together out of the practical situation in the marketplace: thousands of users, thousands of copyright owners, and millions of compositions. Most users want unplanned, rapid, and indemnified access to any and all of the repertoire of compositions, and the owners want a reliable method of collecting for the use of their copyrights. Individual sales transactions in this industry are quite expensive, as would be individual monitoring and enforcement, especially in light of the resources of single composers. Indeed, as both the Court of Appeals and CBS recognize, the costs are prohibitive for licenses with individual radio stations, nightclubs, and restaurants . . . and it was in that milieu that the blanket license arose.

See also WORLD INTELLECTUAL PROPERTY ORGANIZATION, *COLLECTIVE MANAGEMENT OF COPYRIGHT AND RELATED RIGHTS* 130 (2002):

In the case of “performing rights,” reprographic reproduction rights and the rights in respect of simultaneous and unchanged retransmission of broadcast programs, collective administration is an indispensable means of the exercise of the exclusive rights to authorize the uses concerned. . . . The number and circumstances of uses and the number and variety of works used make it practically impossible for the users to identify the right owners in due time, ask for their authorization, negotiate their remuneration and other conditions of the use and to pay the fees on an individual basis.

104. See, e.g., Gervais, *supra* note 70, at 344 (“CMOs offer rightsholders the possibility of carrying a greater weight when negotiating with the larger users.”).

Chinese books for the Google Books Project.”¹⁰⁵ However, Google subsequently found out that CWWCS only had about 6000 members, representing less than 10% of all Chinese authors.¹⁰⁶ In other words, Google could still face liability from the other 90% of Chinese authors after paying the collecting society. Google quickly walked away, as any rational business would do. This story explains why the less representative a collecting society is, the more likely it is to desire an ECL. However, there is an obvious irony: the ECL system presupposes a highly representative and efficient collecting society that operates in a well-functioning copyright market;¹⁰⁷ in reality, the strongest proponents of the ECL system are oftentimes nascent collecting societies that have yet to accumulate sufficient memberships in a market dominated by a large number of foreign or otherwise nonmember works.¹⁰⁸

Nevertheless, the proposed ECL is by no means the only way to increase representation. Another method is for collecting societies to improve their services and attract more members. In this sense, the fundamental concern about the proposed ECL is its potential effect on the incentives for collecting societies to offer authors better services by implementing new technologies and increasing operational efficiency.

Currently, without the benefits the proposed ECL provides, Chinese collecting societies sometimes provide an indemnification clause as part of their license agreements. Accordingly, collecting societies would defend and hold their users harmless from all copyright claims including those from non-members.¹⁰⁹ These collecting societies essentially operate in a way

105. See Ziqiang Wang, *Media Interaction Conference for the Third Amendment to Copyright Law (manuscript)*, SINA (Apr. 25, 2012), http://ent.sina.com.cn/y/2012-04-25/17063615203_7.shtml.

106. See NCAC, *RMB 170 Million Was Distributed in 2011*, GOV.CN (Dec. 4, 2010), http://www.gov.cn/gzdt/2012-12/04/content_2282411.htm.

107. See *supra* note 33 and accompanying text.

108. See, e.g., Riis & Schovsbo, *supra* note 49, at 937 (noting Norwaco, the Norwegian collecting society that manages cable retransmission rights, requested the ECL power even though of the two hundred broadcasting channels in Norway, only fourteen were actually Norwegian); see also Jiang, *supra* note 95, at 732 (2013) (noting the ECL was proposed in China, a country “[w]ith nascent collective societies and a vulnerable copyright regime”); Gervais, *supra* note 15, at 4 (“The advantage of the extended collective license system in Canada would be to place small collective management organizations (CMOs) on the same footing as large CMOs.”).

109. The indemnification clause appears to be common practice in the United States and in Europe. See Koskinen-Olsson, *supra* note 33, at 292; cf. *Columbia Broad. Sys., Inc. v. American Soc. of Composers*, 562 F.2d 130, 140 (2d Cir. 1977):

There is not enough evidence in the present record to compel a finding that the blanket license does not serve a market need for those who wish

similar to liability insurance companies.¹¹⁰ In the United States, prospective users could obtain virtually total freedom to use any musical works by purchasing blanket licenses in a competitive market of collecting societies that includes ASCAP, BMI, and SESAC.¹¹¹ These successful examples demonstrate how market forces may both diminish transaction costs and give rise to dynamic competition toward efficiency.

In response to overwhelming opposition from the Chinese public, the second draft of the Chinese Copyright Reform Bill narrowed the scope of the proposed ECL to two areas: use of musical and audiovisual works in karaoke parlors; and use of literary, musical, artistic, and photographic works by broadcasting organizations.¹¹² This proposal also makes little sense. Both karaoke parlors and music companies constitute commercial undertakings that operate in a highly organized and competitive marketplace. One would be hard-pressed to identify another pair of willing buyers and willing sellers that are more prepared to engage in market negotiations. These sectors arguably need the ECL the least. The ECL power would also be redundant for broadcasters should they continue to receive a compulsory license for uses of copyrighted works in their programs.¹¹³

On a related note, Nordic countries originally designed the ECL not as a mechanism to encroach extensively on the exclusive rights of copyright owners but as a favored alternative to compulsory licenses.¹¹⁴ Should Chinese policymakers be determined to implement the Nordic approach faithfully, they should start by transforming existing compulsory licenses into the ECL regime. First, it would alleviate the rigidity and insensitivity to market reality inherent in existing compulsory licenses because authors

full protection against infringement suits or who, for some other business reason, deem the blanket license desirable. The blanket license includes a practical covenant not to sue for infringement of any ASCAP copyright as well as an indemnification against suits by others.

110. This contractual design is one of the reasons why Chinese collecting societies called for a new limitation on liability provision under copyright law for collecting society licensees. See Article 70, First Draft, *supra* note 4; see also Jichao Ma, *The Necessity and Urgency of Implementing Extended Collective Licenses in Our Country*, SINA (Nov. 30, 2011), http://blog.sina.com.cn/s/blog_593badd10100xdbv.html (describing how CAVCA assisted karaoke parlors in handling lawsuits authors filed).

111. See Music Marketplace Study, *supra* note 34, at 19.

112. Article 60, Second Draft, *supra* note 4.

113. Article 43, Copyright Law of the People's Republic of China of 2010 (effective on Apr. 1, 2010), <http://www.ncac.gov.cn/chinacopyright/contents/479/17542.html> [hereinafter "Chinese Copyright Law"].

114. See Riis & Schovsbo, *supra* note 48, at 473.

would have a choice to opt out of the ECL and instead license their works directly. Second, for authors who prefer to stay within the ECL system, in most cases their copyright royalties would be decided through arms-length negotiations between collecting societies and prospective users. Third, the Chinese government would have a limited number of pilot projects to examine the long-term dynamics of the ECL regime first without unduly disturbing the status quo and vested interests.¹¹⁵

III. COMPULSORY LICENSE

The mechanical license under 17 U.S.C. § 115 originates from the first compulsory license in U.S. history that Congress established in 1909.¹¹⁶ It allows anyone to make and distribute phonorecords of a musical work after the work has been distributed to the public in the United States in phonorecords with permission, provided that the user serves a notice of intention, presents statements of account, and pays government-set royalties on a monthly basis.¹¹⁷ Mechanical licenses have opened the gate for other compulsory licenses to proliferate. Today, compulsory licenses cover usages ranging from non-interactive digital public performances of sound recordings¹¹⁸ and related ephemeral recordings¹¹⁹ to retransmission by satellite and cable of broadcast programs,¹²⁰ over-the-air transmission of musical works by public broadcasting entities,¹²¹ and a public levy imposed on digital audio recording devices and media.¹²² ASCAP and BMI, which are collecting societies that manage public performance rights of musical works, are also subject to price control by the Southern District of New

115. The U.S. government appears to follow this cautious approach regarding the online music market. See Music Marketplace Study, *supra* note 34, at 1.

116. See Howard B. Abrams, *Copyright's First Compulsory License*, 26 SANTA CLARA COMPUT. & HIGH TECH. L.J. 215, 216 (2010).

117. 17 U.S.C. § 115(b)(1) (2012).

118. § 114(f).

119. § 112(e).

120. §§ 111, 119, 122.

121. § 118(b).

122. § 1003.

York rate courts.¹²³ The provisions on compulsory licenses account for approximately 40% of the 280-page U.S. Copyright Act as it stands.¹²⁴

The Chinese Copyright Law currently includes a provision similar to § 115 of the U.S. Copyright Act:

A producer of sound recording may use a musical work that has been legally recorded in a sound recording to produce another sound recording without permission of the copyright owner, provided that she pays statutorily set royalties. Nevertheless, such a use is not permitted for any work for which the copyright owner has declared that the use is prohibited.¹²⁵

The key difference between U.S. and Chinese mechanical licenses is that copyright owners are permitted to opt out of the compulsory license under the Chinese regime. In this sense, the Chinese provision is more analogous to the ECL system than a typical compulsory license. In any event, the provision has not drawn much public attention before the copyright reform because Chinese musicians and music companies routinely opt out.

The first draft of the Chinese Copyright Reform Bill proposed to transform the mechanical license into a compulsory license by removing the opt-out option in the current provision.¹²⁶ The proposed change sparked an outcry in the Chinese music industry. A number of top Chinese musicians publicly voiced their concerns that the compulsory license would be tantamount to an endorsement of piracy and a disincentive to invest in original music compositions.¹²⁷ The Record Industry Commission of the China Audio-Video Association, which mostly represents potential users of the compulsory license, issued a public statement lamenting that the

123. ASCAP and BMI are governed by consent decrees imposed as a result of antitrust complaints filed by the Antitrust Division of the U.S. Department of Justice. It brought a civil action in 1941 which led to the first consent decree (U.S. v. ASCAP, 1940-1943 Trade Cas. (CCH) ¶ 56, 104 (S.D.N.Y. 1941)); it was then modified in 1950 (U.S. v. ASCAP, 1950-1951 Trade Cas. (CCH) ¶ 62, 595 (S.D.N.Y. 1950)), again in 1960 (U.S. v. ASCAP, 1960 Trade Cas. (CCH) ¶ 69, 612 (S.D.N.Y. 1960)), and most recently in 2001 (U.S. v. ASCAP, 2001 WL 1589999 (S.D.N.Y.)). The first consent order regarding BMI was issued in 1941 (U.S. v. BMI, 1940-1943 Trade Cas. (CCH) ¶ 56, 096 (E.D. Wisc. 1941)), and was modified twice: in 1966 (U.S. v. BMI, 1966 Trade Cas. (CCH) ¶ 71, 941 (S.D.N.Y. 1966)) and 1994 (U.S. v. BMI, 1996-1 Trade Cas. (CCH) ¶ 71, 378 (S.D.N.Y. 1994)).

124. The full text of the U.S. Copyright Act is available at <http://copyright.gov/title17/>.

125. Article 40, Chinese Copyright law, *supra* note 113.

126. Article 46, First Draft, *supra* note 4.

127. See Tang Yue, *Songwriters Say Copyright Draft Doesn't Protect*, CHINA DAILY (Apr. 6, 2012), http://www.chinadaily.com.cn/china/2012-04/06/content_14986887.html (quoting Gao Xiaosong and Li Guangping).

proposal would deprive copyright owners of their licensing rights, destroy the business models of the music industry, and exacerbate the difficulties of musicians who have been fighting an uphill battle with widespread copyright piracy.¹²⁸

The NCAC explained that the compulsory license would be useful for preventing musical composition monopolies.¹²⁹ This explanation, according to musicians, showed how much the government officials had lost touch with the reality in the music industry.¹³⁰ Musicians contended that, facing widespread piracy and competing with free illegal copies, most of them struggle to make ends meet. At present, none of the Chinese music companies is even financially strong enough to go public, let alone create a monopoly.¹³¹ By contrast, some downstream users like Baidu, Alibaba, and Tencent, which Chinese policymakers intend to protect from any monopoly, are among the most powerful companies in the world with dominant market positions and multi-billion dollar assets.¹³² A compulsory license that deprives musicians of their exclusive rights would amplify the already massive bargaining imbalance between technology providers and content providers.

The concerns of Chinese musicians and music companies are not unfounded, especially in light of their previous experience with another compulsory license for broadcasting organizations. Under the Chinese Copyright Law of 1990, radio and television stations that broadcast published sound recordings for non-commercial purposes enjoyed a total exemption from any claims of copyright owners, performers, or sound

128. See *Singing Work: Copyright Office Representatives to Ignore the Voice of the Artist Published Appeal Dissatisfied*, CAIJING (Apr. 26, 2012), <http://industry.caijing.com.cn/2012-04-26/111829501.html> (full text of appeal of Record Industry Commission of the Chinese Audio-Video Association).

129. See *Record: The Third Revision of Copyright Law Press Conference*, SINA (Apr. 25, 2012), <http://ent.sina.com.cn/y/2012-04-25/17063615203.shtml> (transcripts of the presentation by Wang Ziqiang, Department of Law and Treaties, NCAC).

130. See Li Yi, *Chinese Musicians Association Convened an Emergency Conference to Discuss the Controversial Clauses of Copyright Law Amendment*, SINA (Apr. 10, 2012), <http://ent.sina.com.cn/c/2012-04-10/11173601191.shtml> (quoting Song Ke); Liu Ren & Jiang Xu, *The Public Enquiries of the Draft Copyright Law Aroused Controversies*, STATE INTELLECTUAL PROPERTY OFFICE OF THE P.R.C. (Apr. 13, 2012), http://www.sipo.gov.cn/mtjj/2012/201310/t20131023_830773.html (quoting Zang Yanbin).

131. Ren & Xu, *supra* note 130.

132. See *Market Capitalization of the Largest Internet Companies Worldwide as of May 2015 (in Billion U.S. Dollars)*, STATISTA (June 2016), <http://www.statista.com/statistics/277483/market-value-of-the-largest-internet-companies-worldwide/>.

recording producers.¹³³ This provision was arguably inconsistent with the Berne Convention¹³⁴ and the three-step test under the TRIPS Agreement.¹³⁵ Therefore, right before China joined the WTO on December 11, 2001,¹³⁶ the Chinese Copyright Law of 2001 turned the exemption into a compulsory license: “Radio and television stations may broadcast published sound recordings without the permission of copyright owners but shall pay remunerations, unless the relevant parties have agreed otherwise. Detailed measures shall be formulated by the State Council.”¹³⁷ However, for the first eight years after the enactment of the compulsory license, Chinese broadcasting organizations, which are mostly resourceful state-owned enterprises, had wielded their lobbying powers to effectively block any initiatives to stipulate the detailed measures.¹³⁸ During that period, they continued to use all musical works for free. It was not until 2010 that radio and television stations gradually started negotiating with the MCSC, the collecting society responsible for managing the compulsory license.¹³⁹ This experience taught the music industry how vulnerable compulsory licenses were to regulatory capture by lobbying groups.

As a result of strong opposition from Chinese musicians and music companies, the third draft of the Chinese Copyright Reform Bill totally removed the compulsory license for mechanical reproductions together with the pre-existing provision.¹⁴⁰ In recent years, compulsory licensing has actually returned to the spotlight in copyright scholarship. It has been proffered as a solution to orphan works and mass digitization.¹⁴¹ Several commentators suggest legalizing peer-to-peer file sharing in exchange for a levy imposed on information technologies including Internet access.¹⁴² The

133. Article 43, Copyright Law of People’s Republic of China of 1990 (effective on Jun. 1, 1991), <http://www.people.com.cn/zixun/flfgk/item/dwjff/falv/7/7-2-01.html>.

134. Berne Convention, *supra* note 30, Article 11 bis(1).

135. TRIPS Agreement, *supra* note 31, Article 13.

136. *See Member Information: China and the WTO*, WORLD TRADE ORG., https://www.wto.org/english/thewto_e/countries_e/china_e.htm.

137. Article 44, Chinese Copyright Law, *supra* note 113.

138. *See Interim Measures for the Payment of Remuneration for Sound Recordings Played by Radio and Television Stations* (effective on Jan. 1, 2010), http://www.gov.cn/zwjk/2009-11/17/content_1466687.htm.

139. MCSC, 2010 ANNUAL REPORT 3, <http://www.mcsc.com.cn/pdf/phpIK0X7C.pdf>.

140. Third Draft, *supra* note 4.

141. *See, e.g.,* Robert Kirk Walker, *Negotiating the Unknown: A Compulsory Licensing Solution to the Orphan Works Problem*, 35 CARDOZO L. REV. 983, 986 (2014).

142. *See* WILLIAM W. FISHER III, PROMISES TO KEEP: TECHNOLOGY, LAW, AND THE FUTURE OF ENTERTAINMENT 199–258 (2004); Peter Eckersley, *Virtual Markets for Virtual Goods: The Mirror Image of Digital Copyright?*, 18 HARV. J.L. & TECH. 85, 100 (2004);

U.S. Supreme Court in *eBay v. MercExchange* reinforces the discretionary powers of federal courts to withhold injunctive relief in intellectual property cases, which is arguably equivalent to a judiciary-imposed compulsory license.¹⁴³ The rejection of mechanical compulsory licenses in Chinese law serves as a reminder to reevaluate the efficacy and efficiency of compulsory licenses in the digital age. This Part focuses on three traditional rationales for compulsory licenses: monopoly prevention, transaction costs, and political compromise.

A. COPYRIGHT AND MONOPOLY

When the 1909 Copyright Act first introduced the compulsory license in the United States, Congress indeed intended to prevent the emergence of a monopoly power.¹⁴⁴ In the early 1900s, when piano rolls became one of the most popular ways for families to enjoy music in private,¹⁴⁵ federal courts faced the thorny question of whether mechanical renderings of musical compositions without authorization constituted copyright infringement.¹⁴⁶ The Supreme Court held in *White-Smith Music Publishing v. Apollo* that a piano roll was not a “copy” within the meaning of the Copyright Act because humans normally were unable to “see and read” the musical composition reproduced on a piano roll.¹⁴⁷

Although the majority decision in the *Apollo* case found piano rolls not liable for copyright infringement, it was accompanied by a powerful concurrence from Justice Holmes pleading for legislative recognition of mechanical rights.¹⁴⁸ Music publishers had been lobbying Congress for a legislative intervention even prior to the *Apollo* decision.¹⁴⁹ Against this

Neil Weinstock Netanel, *Impose a Noncommercial Use Levy to Allow Free Peer-to-Peer File Sharing*, 17 HARV. J.L. & TECH. 1, 35 (2003).

143. See *eBay Inc. v. MercExchange, L.L.C.*, 547 U.S. 388, 392 (2006). For empirical evidence of the overall influences of the eBay decision, see Jiarui Liu, *Copyright Injunctions After eBay: An Empirical Study*, 16 LEWIS & CLARK L. REV. 216, 231 (2012).

144. H.R. REP. NO. 60-2222, at 8 (1909).

145. See GOLDSTEIN, *supra* note 79, at 53; RUSSELL SANJEK, AMERICAN POPULAR MUSIC AND ITS BUSINESS: THE FIRST FOUR HUNDRED YEARS VOLUME II: FROM 1790 TO 1909 383 (1988); BRIAN DOLAN, INVENTING ENTERTAINMENT: THE PLAYER PIANO AND THE ORIGINS OF AN AMERICAN MUSICAL INDUSTRY 53 (2009).

146. See *White-Smith Music Publ'g Co. v. Apollo Co.*, 209 U.S. 1 (1908).

147. *Id.* at 17.

148. *Id.* at 19.

149. See, e.g., S. 6330, 59th Cong. § 1(g) (1906); H.R. 19853, 59th Cong. § 1(g) (1906); see Harry G. Henn, *The Compulsory License Provisions of the U.S. Copyright Law*, COPYRIGHT LAW REVISION, SENATE COMM. ON THE JUDICIARY, 86TH CONG., 1ST SESS., STUDIES PREPARED FOR THE SUBCOMM. ON PATENTS, TRADEMARKS, AND COPYRIGHTS OF THE COMM. ON THE JUDICIARY 3 (COMM. PRINT 1960) (STUDY NO. 5).

backdrop, the Aeolian Company set out to actively acquire the not-yet-existing mechanical rights in musical compositions speculating that Congress would soon pass a bill to totally reverse the *Apollo* holding.¹⁵⁰ The Aeolian Company was a leading manufacturer of piano rolls and received a patent on its player piano, the Pianola.¹⁵¹ Due to the legal uncertainty surrounding mechanical rights, the Aeolian Company managed to quickly accumulate a significant share in the music market on relatively favorable terms and conditions: Eighty-seven members of the Music Publishers Association granted exclusive mechanical licenses, which in the aggregate represented 381,598 compositions and accounted for 43% of the total market.¹⁵²

The anticompetitive implications of the above practice deeply concerned Congress in 1909, which warned that, if mechanical rights were formally enacted, “not only would there be a possibility of a great music trust in this country and abroad, but arrangements are being actively made to bring it about.”¹⁵³ As a result, the 1909 Copyright Act on the one hand overruled the *Apollo* decision by, for the first time, granting authors exclusive rights to reproduce musical works in phonorecords; on the other hand though, the Act created a compulsory license for subsequent mechanical reproductions of musical works to address the antitrust concern.¹⁵⁴

No matter how realistic the threat of a music cartel was in light of the actions taken by the Aeolian Company (which ceased to exist decades ago), the antitrust rationale for compulsory licenses finds little support in modern

150. See SANJEK, *supra* note 145, at 23, 400.

151. *Id.*; U.S. Patent No. 765,645 (filed Nov. 16, 1899).

152. See SANJEK, *supra* note 145, at 23. The Aeolian Company agreed to pay 10% of the retail prices of piano rolls sold. The other 117 music publishers in the Music Publishers Association controlled 503,597 compositions. Remarkably, iHeart Media—the largest broadcaster in the United States—used the same strategy in 2013 when it entered into licensing agreements with Warner Music Group and a number of independent labels (including Big Machine Records, which represents Taylor Swift) covering both digital performance rights and not-yet-existing terrestrial performance rights of sound recordings. See Ed Christman, *Here’s Why Warner Music’s Deal with Clear Channel Could be Groundbreaking for the Future of the U.S. Music Biz (Analysis)*, BILLBOARD (Sept. 12, 2013), <http://www.billboard.com/biz/articles/news/5694973/heres-why-warner-musics-deal-with-clearchannel-could-be-groundbreaking>. In doing so, iHeartMedia was able to obtain favorable percentage-based rates, unlike the current per-play rates set by Copyright Royalty Judges. See Ben Sisario, *Clear Channel Warner Music Deal Rewrites the Rules on Royalties*, N.Y. TIMES (Sept. 12, 2013), <http://www.nytimes.com/2013/09/13/business/media/clear-channelwarner-music-deal-rewrites-the-rules-on-royalties.html>.

153. H.R. REP. NO. 60-2222, at 8 (1909); S. REP. NO. 60-1108, at 8 (1909).

154. Copyright Act of 1909, Pub. L. No. 60-349, §1(e), 35 Stat. 1075, 1075–76 (1909).

copyright regimes.¹⁵⁵ Copyrights rarely confer market power in the same manner patents do because different copyrighted works are often good (albeit not perfect) substitutes for each other.¹⁵⁶ The high degree of substitutability lies primarily in several legal doctrines in copyright law. First, the idea/expression dichotomy mandates that copyright protection only extend to expressions rather than to ideas in a work of authorship.¹⁵⁷ This principle suggests that a subsequent author could intentionally imitate a pre-existing work as closely as possible, provided the borrowing is limited to unprotected ideas. Second, in accordance with the copying requirement,¹⁵⁸ the exclusive rights of a copyright owner only cover actual copying of her expression. A work of authorship created independently, however similar to a pre-existing one, does not constitute copyright infringement.¹⁵⁹ Indeed, such a work would likely be considered original and entitled to a copyright separate from the pre-existing one.¹⁶⁰ As a result, similar works of authorship abound in the marketplace due to either deliberate imitation or coincidental repetitiveness.

Additionally, the overall concentration in the music market does not reach a level that would generate real market powers for any firms to restrict music production and inflate music prices, at least according to the Federal Trade Commission, which approved the acquisition of EMI by its direct

155. As mentioned above, the majority of music market at the time was controlled by music publishers unrelated to the Aeolian Company. *See supra* note 152 and accompanying text.

156. *See* Edmund W. Kitch, *Elementary and Persistent Errors in the Economic Analysis of Intellectual Property*, 53 VAND. L. REV. 1727, 1730 (2000) (arguing that “copyrights do not prevent competitors from creating works with the same functional characteristics”); GOLDSTEIN, *supra* note 29, at 86 (“Although we would prefer not to admit it, one author’s expression will always be substitutable for another’s.”).

157. 17 U.S.C. § 102(a)–(b) (2012). It is not an overstatement that most countries recognize the idea/expression dichotomy since the TRIPS Agreement includes such a provision. *See* TRIPS Agreement, *supra* note 31, Article 9(2) (“Copyright protection shall extend to expressions and not to ideas, procedures, methods of operation or mathematical concepts as such.”).

158. In other words, the social costs of copyrights are limited access to a work created by the author, while the social costs of patents are limited access to certain inventions created either by the patentee or by any third party.

159. *See, e.g.*, *Sheldon v. Metro-Goldwyn Pictures Corp.*, 81 F.2d 49, 54 (2d Cir. 1936) (“[B]ut if by some magic a man who had never known it were to compose anew Keats’s Ode on a Grecian Urn, he would be an ‘author,’ and, if he copyrighted it, others might not copy that poem, though they might of course copy Keats’s.”).

160. To this extent, copyright law drastically differs from patent law, under which the exclusivity of patent rights is relatively strong, covering not only unscrupulous copying but also independent creation of the same invention.

competitors in 2012.¹⁶¹ In terms of music publishing, three major firms, Sony/ATV Music Publishing (“Sony/ATV”), Universal Music Publishing Group (“UMPG”), and Warner/Chappell Music respectively control 22.4%, 28.9%, and 17.4% of the music publishing market.¹⁶² Because digital technology has lowered the market entry barriers regarding music production and distribution costs, the majors are faced with increasing competition from thousands of independent music publishers including Kobalt Music Group and BMG Chrysalis, which have in the aggregate grown from 31.6% of the market in 2007 to 35% in 2014.¹⁶³

The recording sector, mostly downstream users of mechanical licenses, features a market structure similar to the publishing sector. There are three major labels, Universal Music Group (“UMG”), Sony Music Entertainment, Inc. (“SME”), and Warner Music Group (“WMG”) that respectively hold 27.5%, 22%, and 14.6% market shares.¹⁶⁴ Hundreds of independent labels combined account for 35.1% of record industry revenues. Furthermore, major music labels and major music publishers are subject to common corporate ownership. UMPG is owned by UMG, the Sony Corporation owns SME and half of Sony/ATV, and Warner/Chappell Music is a division of WMG.¹⁶⁵ These factors suggest that music publishers would rarely be able to leverage their bargaining powers against music labels in negotiating mechanical licenses, regardless of any constraints created by a compulsory license.

161. See *FTC Closes Its Investigation into Vivendi, S.A.’s Proposed Acquisition of EMI Recorded Music*, FTC (Sept. 21, 2012), <https://www.ftc.gov/news-events/press-releases/2012/09/ftc-closes-its-investigation-vivendi-sas-proposed-acquisition-emi>; (FTC press release); *FTC Closes Its Investigation Into Sony/ATV Music Publishing’s Proposed Acquisition of EMI Music Publishing*, FTC (June 29, 2012), <https://www.ftc.gov/news-events/press-releases/2012/06/ftc-closes-its-investigation-sonyatv-music-publishings-proposed> (FTC press release).

162. See *Revenue Market Share of the Largest Music Publishers Worldwide from 2007 to 2016*, STATISTA (Feb. 2017), <http://www.statista.com/statistics/272520/market-share-of-the-largest-music-publishers-worldwide/>.

163. See *id.*

164. See Ed Christman, *Music in 2014: Taylor Takes the Year, Republic Records on Top, Streaming to the Rescue*, BILLBOARD (Jan. 9, 2015), <http://www.billboard.com/articles/business/6436399/nielsen-music-soundscan-2014-taylor-swift-republic-records-streaming?page=0%2C2>.

165. See Sebastian Torrelio, *Jody Gerson Appointed Chairman and CEO of Universal Music Publishing Group*, VARIETY (Aug. 1, 2014), <http://variety.com/2014/biz/news/jody-gerson-appointed-chairman-and-ceo-of-universal-music-publishing-group-1201273829>; *Profile: Sony Corp.*, REUTERS, <http://www.reuters.com/finance/stocks/companyProfile?symbol=SNE.N>; *About Us*, WARNER/CHAPPELL MUSIC, <http://www.warnerchappell.com/about>.

Music publishers also have to deal with even more formidable market players of the caliber of Apple, Amazon, and Google in the digital environment. The digital music market is clearly more concentrated than the music publishing market. For instance, iTunes (64%) and Amazon MP3 (16%) combined control 80% of the music download market,¹⁶⁶ and YouTube alone accounts for 77.6% of Internet video visits.¹⁶⁷ Artificially limiting the bargaining powers of music publishers would effectively reinforce the dominant positions of Internet service providers.¹⁶⁸

There is no guarantee that anticompetitive behaviors would not emerge in any copyright industry. However, antitrust statutes incorporate more comprehensive and sophisticated mechanisms than does the Copyright Act to handle anticompetitive concerns, allowing for both specialized government agency and private enforcement.¹⁶⁹ It appears grossly disproportionate to subject a whole industry to a compulsory license for the purpose of redressing a limited number of wrongdoings.

B. TRANSACTION COSTS AND RENT SEEKING COSTS

Another conventional justification for compulsory licenses is minimal transaction costs because prospective users may unilaterally decide to use copyrighted works upon payment of statutory royalties without negotiation with copyright owners.¹⁷⁰ However, recent empirical studies present strong

166. See *Paid Digital Music Download Market in the United States as of September 2012*, STATISTA (Sept. 2012), <http://www.statista.com/statistics/248995/us-paid-music-download-market-distribution/>.

167. See *Leading Multimedia Websites in the United States in October 2016, Based on Market Share of Visits*, STATISTA (Nov. 2016), <http://www.statista.com/statistics/266201/us-market-share-of-leading-internet-video-portals/>.

168. One may argue that a compulsory license could help alleviate the problem of double marginalization resulting from monopolistic complementary markets in information products and information technologies. But this would still not explain why the government should impose a price control on music publishers rather than Internet service providers if the price control is needed at all.

169. See, e.g., *Radio Music License Comm., Inc. v. SESAC*, 29 F. Supp. 3d 487 (E.D. Pa. 2014); *Meredith Corp. v. SESAC, L.L.C.*, 87 F. Supp. 3d 650, 2015 WL 728026 (S.D.N.Y. 2015).

170. See, e.g., H.R. REP. NO. 94-1476, at 89 (1976), *reprinted in* 1976 U.S.C.C.A.N. 5659, 5703 (“[I]t would be impractical and unduly burdensome to require every cable system to negotiate with every copyright owner whose work was retransmitted by a cable system.”); *NBC v. Copyright Royalty Tribunal*, 848 F.2d 1289, 1291 (D.C. Cir. 1988) (“The purpose of this regulatory structure is to facilitate the exploitation of copyrighted materials by removing the prohibitive transaction costs that would attend direct negotiations between cable operators and copyright holders, while at the same time assuring copyright holders compensation for the use of their property.”); AL KOHN & BOB KOHN, *KOHN ON MUSIC LICENSING* 772 (2010) (“[T]he compulsory license has served to

evidence demonstrating that private parties continue to actively bargain in the shadow of compulsory licenses and other liability rules.¹⁷¹

Ironically, a typical example is precisely the first compulsory license. Although the U.S. government has regularly set royalty rates for various mechanical licenses, a majority of music labels approach private clearinghouses, such as the Harry Fox Agency, to negotiate for mechanical licenses.¹⁷² The music labels apparently find the monthly accounting obligations under § 115 too cumbersome and prefer to directly negotiate with the Harry Fox Agency, which offers quarterly accounting as a standard practice.¹⁷³ In other words, the administrative costs incurred for compulsory licenses outweigh the transaction costs involved in private negotiations.

Not only do private parties routinely contract around compulsory license terms, but the U.S. Copyright Act also contains built-in mechanisms to encourage private negotiations prior to, and during the course of, governmental rate-setting proceedings. For instance, Congress requires that voluntary licenses negotiated between copyright owners and prospective users take precedence over the rates and terms set by government agencies.¹⁷⁴ Congress also provides antitrust exemptions for both parties to compulsory licenses so that they could respectively designate common agents to negotiate royalty rates.¹⁷⁵ The rate-setting proceedings specifically set aside a three-week period for settlement negotiations.¹⁷⁶

While compulsory licenses have limited effects in minimizing transaction costs, they usually generate additional administrative costs unseen in private transactions. Both copyright owners and prospective users have strong incentives to invest substantial resources in lobbying

simplify the process of obtaining mechanical licenses and has reduced a significant amount of unnecessary transaction costs.”).

171. See Lemley, *supra* note 24, at 463.

172. See, e.g., *Music Licensing Reform Before the Subcomm. on Intellectual Property of the S. Comm. on the Judiciary*, 109th Cong., 1st Sess. (2005) (statement of Marybeth Peters, Register of Copyrights), <http://www.copyright.gov/docs/regstat071205.html> (“[T]he use of the [compulsory] license appears to have again become almost non-existent; up to this day, the Copyright Office receives very few notices of intention.”); Mechanical and Digital Phonorecord Delivery Rate Determination Proceeding, 74 Fed. Reg. 4509, 4510, 4511, 4520 n.32 (Jan. 26, 2009) (codified at 37 C.F.R. pt. 385) (citing testimony of RIAA expert economist Dr. Steven Wildman) (“As witnesses for both record companies and music publishers have explained, essentially no one uses the compulsory license process—licenses for mechanical royalties . . . are negotiated in the market on a voluntary basis.”).

173. See Loren, *supra* note 25, at 682 n.38.

174. See 17 U.S.C. §§ 112(e)(5), 114(f)(3), 115(c)(3)(E)(i) (2012).

175. See §§ 112(e)(2), 114(e)(1), 115(c)(3)(B).

176. § 803(b)(6)(C)(x).

government agencies¹⁷⁷ for favorable royalty rates (often referred to as rent-seeking costs or influence costs).¹⁷⁸ The legislative history of the 1976 Copyright Act pointed to the sheer magnitude of expenditures by lobbyists trying to influence the licensing rates. In the first comprehensive amendment of the U.S. Copyright Act in sixty years, over 32% of the 1976 hearings record was devoted to rates.¹⁷⁹ ASCAP, which is subject to rate-setting proceedings by rate courts, more recently admitted, “ASCAP and applicants have collectively expended well in excess of one hundred million dollars on litigation expenses related to rate court proceedings, much of that incurred since only 2009.”¹⁸⁰

The proceedings regarding webcaster royalties provide another dramatic illustration.¹⁸¹ While the Digital Performance Right in Sound Recordings Act of 1995 recognizes “the exclusive right to perform the copyrighted work publicly by means of a digital audio transmission,” it simultaneously provides for a compulsory license covering non-interactive webcasting services.¹⁸² However, no royalty negotiations for such services started until the passage of the Digital Millennium Copyright Act of 1998,¹⁸³ which further clarified the rate-setting standards. A large number of webcasters had emerged operating practically in a royalty-free environment by the time the Copyright Arbitration Royalty Panel (“CARP”) proceedings officially concluded in 2002.¹⁸⁴ The Library of

177. Congress created the Copyright Royalty Tribunal (“CRT”), with five commissioners appointed by the President, to adjust the royalty rate. *See* Copyright Act of 1976, Pub. L. No. 94-553, §§ 801–802, 90 Stat. 2541, 2594–96 (1976). The CRT was replaced in 1993 by the Copyright Arbitration Royalty Panel (“CARP”) system; rather than permanent appointees, the CARP arbitrators were convened for specific rate proceedings. *See* Copyright Royalty Tribunal Reform Act of 1993, Pub. No. 103-198, §802, 107 Stat. 2304, 2305 (1993). The CARP system, in turn, was replaced in 2004 by the current system, the Copyright Royalty Board (“CRB”), which is composed of three administrative judges (“Copyright Royalty Judges”) appointed by the Librarian of Congress. *See* 17 U.S.C. § 801 (2012); Copyright Royalty and Distribution Reform Act of 2004, Pub. L. No. 108-419, 118 Stat. 2341 (2004).

178. *See* Paul Milgrom & D. J. Roberts, *Bargaining Costs, Influence Costs, and the Organization of Economic Activity*, in PERSPECTIVES ON POSITIVE POLITICAL ECONOMY 57–89 (James E. Alt & Kenneth A. Shepsle eds., 1990).

179. *Copyright Law Revision: Hearings on H.R. 2223 Before the Subcomm. on Courts, Civil Liberties and the Administration of Justice of the House Comm. on the Judiciary*, 94th Cong., 1st Sess., pts. 1–3 (1976).

180. *See* Music Marketplace Study, *supra* note 34, at 93.

181. 17 U.S.C. § 114(f) (2012).

182. Pub. L. No. 104-39, §4, 109 Stat. 336 (1995).

183. Pub. L. No. 105-304, tit. IV, 112 Stat. 2860 (1998).

184. Determination of Reasonable Rates and Terms for the Digital Performance of Sound Recordings and Ephemeral Recordings, 67 Fed. Reg. 45,240, 45,272 (July 8, 2002).

Congress (“LOC”) affirmed the CARP decision to set per-performance rates, including \$0.0007 per performance for commercial webcasters and \$0.0002 per performance for non-commercial webcasters.¹⁸⁵ Webcasters who felt the CARP rates overly burdensome acted immediately to lobby for an amendment in Congress, which passed the Small Webcasters Settlement Act of 2002 to suspend the implementation of the CARP decision and encourage SoundExchange to negotiate directly with webcasters for alternative royalty structures.¹⁸⁶ Both parties eventually reached agreements (effective through 2005), under which small webcasters would pay a graduated percentage-of-revenue rate and non-commercial webcasters would pay a flat rate.¹⁸⁷ In essence, the CARP proceedings, which reportedly incurred \$25 million in legal fees and witness costs, were completely superseded by private negotiations.¹⁸⁸

The drama repeated itself five years later despite Congress attempted to streamline rate-setting procedures by replacing the ad hoc CARP with the Copyright Royalty Board (CRB), consisting of three standing Copyright Royalty Judges.¹⁸⁹ In 2007, the CRB similarly ruled in favor of a per-performance structure for commercial webcasters, establishing a graduated rate starting from \$0.0008 for 2006.¹⁹⁰ Several webcasters expectedly objected to the CRB rates and filed appeals in the D.C. Circuit.¹⁹¹ Meanwhile, they quickly returned to Congress, which again enacted legislative solutions following the exact pattern of the Small Webcaster Settlement Act of 2002.¹⁹² With the authority from Congress to engage in private negotiations for alternative licensing rates, SoundExchange reached another round of agreements with a wide variety of webcasters (effective through 2015). As a result, commercial pureplay internet radio services

185. *See id.*

186. *See generally* Pub. L. No. 107-321, 116 Stat. 2780 (2002).

187. Notification of Agreements under the Small Webcaster Settlement Act of 2002, 67 Fed. Reg. 78510 (Dec. 24, 2002); Notification of Agreements under the Webcaster Settlement Act of 2002, 68 Fed. Reg. 35008 (June 11, 2003).

188. *Copyright Royalty and Distribution Reform Act of 2003: Hearing Before the Subcomm. on Courts, The Internet, and Intellectual Property of the House Comm. on the Judiciary*, 108th Cong., 1st Sess. 28, 31 (2003) (statement of Michael J. Remington, former staff counsel, IP Subcommittee, House Committee on the Judiciary).

189. *See* Copyright Royalty and Distribution Reform Act of 2004, Pub. L. No. 108-419, 118 Stat. 2341 (2004).

190. 37 C.F.R. § 380.3(a)(1) (2012).

191. *See* Terry Hart, *A Brief History of Webcaster Royalties*, COPYHYPE (Nov. 28, 2012), <http://www.copyhype.com/2012/11/a-brief-history-of-webcaster-royalties/>.

192. *See* Webcaster Settlement Act of 2008, Pub. L. No. 110-435, 122 Stat. 4974 (2008); Webcaster Settlement Act of 2009, Pub. L. No. 111-36, 123 Stat. 1926 (2009) (extending the timeframe for negotiation).

agreed to pay 25% of gross revenues or a per-performance rate lower than the CRB rate, whichever greater.¹⁹³ The clock for the next round of CRB proceedings has started ticking as the rate-setting under Section 114 takes place every five years by operation of law.¹⁹⁴

The above examples reveal that private negotiation and government lobbying are normally intertwined in the context of compulsory licenses. Contrary to the conventional wisdom that compulsory licenses simplify licensing processes, they essentially add rent-seeking costs on top of transaction costs.

C. POLITICAL COMPROMISE

Monopoly concerns and transaction costs are not the whole story for traditional compulsory licenses. In cases involving new information technologies, Congress has sometimes enacted a compulsory license to broker a compromise between dueling stakeholder groups of technology providers who demanded unabridged access to copyrighted works and of copyright owners who asserted exclusive rights over new uses of their works.

Take the *Apollo* case as an example. Piano rolls, a technological innovation back then, produced new uses of musical compositions, and so music publishers approached the judiciary for legal remedies by suing the technology manufacturer.¹⁹⁵ The district court in *Apollo* ruled in favor of the defendant due to the lack of clear guidance in the Copyright Act.¹⁹⁶ By the time the case was finally before the Supreme Court several years later, the new uses had become sufficiently widespread that the Supreme Court had to reluctantly affirm the district court's decision and call upon Congress to reform the outdated statute in response to new technologies.¹⁹⁷ Faced with forceful lobbying by a thriving new technology industry and the potential public relations nightmare of outlawing everyday activities by millions of consumers, Congress was compelled to mediate a political compromise through a compulsory license:¹⁹⁸ Technology providers could continue their business models without the constraints of exclusive rights;

193. Notification of Agreements under the Webcaster Settlement Act of 2009, 74 Fed. Reg. 34,796, 34,799–800 (July 17, 2009).

194. 17 U.S.C. § 804(b)(3) (2012).

195. *White-Smith Music Publishing Co. v. Apollo Co.*, 209 U.S. 1, 8 (1908).

196. *Id.* at 8.

197. *Id.* at 9 (indicating millions of piano rolls and player pianos were manufactured).

198. 17 U.S.C. § 115 (2012).

in exchange, copyright owners received presumably equitable remuneration to maintain creative incentive.¹⁹⁹

Such political dynamics became a recurrent theme for other compulsory licenses. In the 1960s and 70s, when copyright owners brought actions against cable operators for augmenting local signals and importing distant signals, the Supreme Court twice affirmed district court decisions in favor of the defendants, holding that cable systems did not involve public performance because they were more aligned with a viewer function than a broadcaster function.²⁰⁰ Two years later, in the 1976 Copyright Act, Congress responded by carving out a compulsory license for cable retransmission.²⁰¹ In the 1980s, while movie studios sought injunctive relief to enjoin the distribution of analog video recorders, the Supreme Court again sided with the district court, which found no indirect liabilities for a manufacturer whose machines facilitated time shifting by television viewers.²⁰² Seven years later, the advent of digital audio recorders (which could generate unlimited copies without quality degradation) started to push the envelope of the *Sony* holding.²⁰³ Congress consequently enacted the Audio Home Recording Act of 1992 to impose a levy on digital audio recording devices and media, and prohibited any copyright actions against the providers and their consumers.²⁰⁴

By contrast, the market landscape would have been dramatically different if a court initially did not hesitate to uphold the exclusive rights of copyright owners over new uses of copyrighted works. Taking peer-to-peer file sharing as an example, when a district court issued a preliminary injunction against Napster in 1999,²⁰⁵ which was essentially affirmed by the Ninth Circuit in 2002,²⁰⁶ music labels and technology providers actively engaged in private negotiations for voluntary licenses, both inside and

199. See, e.g., Jane C. Ginsburg, *Fair Use for Free, or Permitted-but-Paid?*, 29 BERKELEY TECH. L. J. 1383, 1386 (2014); Robert P. Merges, *Compulsory Licensing vs. the Three "Golden Oldies": Property Rights, Contracts, and Markets*, POL'Y ANALYSIS NO. 508 (Cato Inst., Wash., D.C.), Jan. 15, 2004, at 1; Tim Wu, *Copyright's Communications Policy*, 103 MICH. L. REV. 278, 290, 311 (2004).

200. *Teleprompter Corp. v. CBS*, 415 U.S. 394, 408–10 (1974); *Fortnightly Corp. v. United Artists Television, Inc.*, 392 U.S. 390, 399–401 (1968).

201. 17 U.S.C. § 111 (2012).

202. *Sony Corp. of Am. v. Universal City Studios, Inc.*, 464 U.S. 417, 420–21 (1984).

203. *Cahn v. Sony Corp.*, No. 90 Civ. 4537 (S.D.N.Y. July 11, 1991).

204. Pub. L. No. 102-563, 106 Stat. 4237 (1992) (codified as amended at 17 U.S.C. §§ 1001–1010).

205. *A&M Records, Inc. v. Napster, Inc.*, 114 F. Supp. 2d 896, 901 (N.D. Cal. 2001), *aff'd in part and rev'd in part*, 239 F.3d 1004 (9th Cir. 2001).

206. *A&M Records, Inc. v. Napster, Inc.*, 239 F.3d 1004 (9th Cir. 2001).

outside of the legal proceedings.²⁰⁷ A robust digital music market has quickly emerged, starting from the launch of the iTunes store in 2003, which not only changed the fate of then-ailing Apple but also permanently changed the way people purchase music.²⁰⁸ Unsurprisingly, Congress has not issued any compulsory license for file sharing despite the pleadings from several high-profile scholars.²⁰⁹

A compulsory license that substitutes reasonable remuneration for exclusive rights might not obstruct a copyright market if the government-set rates could accurately reflect the market values that would otherwise be revealed through private transactions. However, empirical evidence from several natural experiments appears to suggest that the government has an inherent tendency to underestimate the value of copyrighted works in rate-setting proceedings. For instance, no matter how closely the two-cent rate for mechanical licenses imitated the market rates in 1909, it would be hard to justify that the rate stayed two cents for almost seventy years from 1909 to 1978.²¹⁰ During the same period of time, overall inflation increased by 500%, suggesting the real value of the two-cent rate decreased by 500%.²¹¹ Likewise, the current 9.1-cent rate in 2015 remains significantly lower than the two-cent rate in 1909, which amounts to over fifty cents in 2015 after adjustment for inflation.

The rapid depreciation of statutory royalties dramatically contrasts the growing importance of musical works in the digital age.²¹² In 2004, several music publishers and music labels, uncertain whether ringtones fell into the definition of Digital Phonorecord Delivery under the compulsory license,

207. See Jeff Leeds, *Bertelsmann Reaches Deal with EMI over Napster*, N.Y. TIMES (Mar. 27, 2007), http://www.nytimes.com/2007/03/27/business/media/27music.html?_r=0 (reporting Bertelsmann reached an agreement to invest in Napster).

208. See Press Release, Apple, iTunes Store Tops 10 Billion Songs Sold (Feb. 25, 2010), <http://www.apple.com/pr/library/2010/02/25itunes.html>.

209. See *supra* note 142 and accompanying text.

210. See U.S. BUREAU OF LABOR STATISTICS, *Consumer Price Index*, <http://www.bls.gov/cpi/> (documenting the general inflation during the same period of time).

211. See U.S. COPYRIGHT OFFICE, *Mechanical License Royalty Rates*, <http://www.copyright.gov/licensing/m200a.pdf>. Congress raised the rate in the 1976 Copyright Act, which became effective in 1978.

212. It has been reported that, in 2014, there were 434,695,663,626 total music streams across Spotify, YouTube, Vevo, Soundcloud, Vimeo, and Rdio. This reflects a 95% increase from 2013 and a 363% increase from 2012. See Zach O'Malley Greenburg, *Truth in Numbers: Six Music Industry Takeaways from Year-End Data*, FORBES (Jan. 22, 2015), <http://www.forbes.com/sites/zackomalleygreenburg/2015/01/22/truth-in-numbers-six-music-industry-takeaways-from-year-end-data/>; NEXT BIG SOUND, *The State of The Industry*, <https://www.nextbigsound.com/industryreport/2014>.

entered into voluntary licensing agreements.²¹³ These market transactions established a royalty rate for ringtones at twenty-four cents per use,²¹⁴ drastically higher than the 9.1-cent rate set by the government. Remarkably, ringtones usually use short excerpts of musical works, which highlights how much the government-set rate undervalues full-length works.²¹⁵ Similarly, several music publishers sought in 2013 to withdraw new media rights from ASCAP and BMI to bypass the rate-setting proceedings at rate courts and negotiate directly with service providers for higher royalties.²¹⁶ As a result, EMI received a rate equivalent to the ASCAP rate of 1.85% for non-interactive streaming services but without a deduction for ASCAP surcharges; Sony/ATV obtained a prorated share of 5% (equivalent to a 2.28% implied rate for ASCAP); and UMG negotiated a prorated share of 7.5% (equivalent to a 3.42% implied rate for ASCAP).²¹⁷

Several phenomena in the copyright market explain why compulsory licenses may systematically undercompensate authors for the market values of copyright licenses. First, the unauthorized version of a copyrighted work, if distributed to the public in a poor quality, inflicts a reputational harm to the author, which may scarcely be quantifiable in any compulsory licenses.²¹⁸ Such reputational harms are actually as relevant to financial

213. Mechanical and Digital Phonorecord Delivery Rate Adjustment Proceeding, 71 Fed. Reg. 64,303, 64,308–09 (Nov. 1, 2006).

214. 37 C.F.R. § 385.3(b) (2015).

215. Mechanical and Digital Phonorecord Delivery Rate Determination Proceeding, 74 Fed. Reg. 4509, 4522 (Jan. 26, 2009) (Music publishers introduced the negotiated agreements as marketplace benchmarks and secured a much higher rate for ringtones than the rate for full songs.).

216. Pandora Media, Inc. v. Am. Soc’y of Composers, Authors, & Publishers, 6 F. Supp. 3d 317, 330, 339 – 40, 355 (S.D.N.Y. 2014).

217. These negotiated rates were later overruled by the rate courts. See *In re* Petition of Pandora Media Inc., No. 12 Civ. 8035(DLC), 2013 WL 5211927, at *11 (S.D.N.Y. Sept. 17, 2013); Broadcast Music, Inc. v. Pandora Media, Inc., No. 13 Civ. 4037(LLS), 2013 WL 6697788, at *5 (S.D.N.Y. Dec. 19, 2013).

218. See, e.g., Nintendo of Am. Inc. v. Lewis Galoob Toys, Inc., No. 90-15936, 1991 WL 5171, at *3 (9th Cir. Jan. 24, 1991) (“A large portion of the harm to Galoob and Nintendo from either granting or denying the preliminary injunction would be financially compensable. However, Nintendo introduced evidence in the court below that the sale of Game Genie could cause irreparable harm to its design strategy, reputation, and its right to create derivative works.”); Columbus Rose Ltd. v. New Millennium Press, No. 02 CIV. 2634(JGK), 2002 WL 1033560, at *11 (S.D.N.Y. May 20, 2002) (holding that plaintiff’s “name and artistic reputation are his major assets,” which “cannot be remedied by a monetary award”) (internal quotation marks omitted); Art Line, Inc. v. Universal Design Collections, Inc., 966 F. Supp. 737, 744–45 (N.D. Ill. 1997); Clifford Ross Co. v. Nelvana, Ltd., 710 F. Supp. 517, 520–21 (S.D.N.Y. 1989) (“[I]t is well established that loss to artistic reputation . . . cannot be compensated for in money damages.”), *aff’d*, 883 F.2d 1022 (2d Cir. 1989).

interests as to moral rights because “the ultimate commercial success of an ‘artist’ often depends on name recognition and reputation with the value and popularity of each succeeding work depending upon the ‘name’ established through commercial exploitation of preceding works.”²¹⁹ In voluntary agreements, copyright owners may insist on extensive involvement in the editing process, request a prior approval for every proposed change, or otherwise wield quality control over the licensed versions. It is unclear whether such quality control is practical, or even possible, in compulsory licenses. Second, creative industries involve a notoriously high degree of uncertainty in consumer preferences. Therefore, copyright owners traditionally maintain a large portfolio of different works that they use to cross-subsidize experimental or pioneer works with lucrative revenues from bestsellers.²²⁰ Uniform reasonable royalties set for compulsory licenses would likely undercompensate copyright owners for their business risks in producing a variety of works. Third, copyright owners would lose the ability to grant an exclusive license in the shadow of a compulsory license. The total royalties from several nonexclusive licenses would probably still be less than those from a single exclusive license due to the erosion of market power by multiple competitors. Fourth, unlike private negotiations, the rate-setting proceedings are susceptible to lobbying pressures in a political environment where lower royalty rates are understandably more popular among potential voters and powerful technology sectors.²²¹ Fifth, compulsory licenses may result in uneven bargaining positions between copyright owners and prospective users. Where the government-set rate is too high, a licensee could always ask for a discounted rate from a copyright owner or simply cease using her work.²²² By contrast, if the government-set rate is too low, a copyright owner would be unable to stop the other party from using her work and therefore have much less leverage to bargain for a higher rate.²²³ In most cases, the compulsory license creates a ceiling for the

219. *Concrete Mach. Co. v. Classic Lawn Ornaments, Inc.*, 843 F.2d 600, 611 (1st Cir. 1988); *see also Ty, Inc. v. GMA Accessories, Inc.*, 132 F.3d 1167, 1173 (7th Cir. 1997); *Gilliam v. Am. Broad. Cos.*, 538 F.2d 14 (2d Cir. 1976).

220. *See* GOLDSTEIN, *supra* note 29, at 83; Barry W. Tyerman, *The Economic Rationale for Copyright Protection for Published Books: A Reply to Professor Breyer*, 18 UCLA L. REV. 1100, 1121 (1971).

221. *See supra* note 184 and accompanying text.

222. The dramas surrounding webcasting royalty rates somewhat illustrate the bargaining power of copyright users to demand lower rates under compulsory licenses. *See supra* text accompanying note 181.

223. However, it does not follow that a copyright owner would never have any chance to bargain for higher royalties in the context of a compulsory license. For narrow exceptions, *see* Lemley, *supra* note 24, at 463.

amount copyright owners may realistically bargain for in private negotiations.²²⁴ This explains why both parties often incur substantial expenses to litigate and lobby for a compulsory license they rarely use in practice.²²⁵

IV. ORPHAN WORKS AND MASS DIGITIZATION

“Orphan works” refer to the situations where a prospective user who wishes to obtain a license for a copyrighted work cannot identify or locate the copyright owner with a reasonably diligent search.²²⁶ Faced with the possibility of expensive copyright litigation and ensuing statutory damages, risk-averse users such as libraries, museums, and archives would often shy away from a projected use that could otherwise generate substantial social value. Meanwhile, the copyright owner, if located, might be more than happy to grant the user a license. As a result, excessive transaction costs involved in locating relevant parties in effect prevent a mutually beneficial transaction from taking place.

The increased attention to orphan works stems from a number of modern developments in international copyright regimes. First, many countries, including the United States and most European nations, have extended the duration of copyright protection to life plus seventy years.²²⁷ Generally, the longer the time having elapsed since a copyrighted work was created and published, the more difficult it is to ascertain its current ownership. Second, the Berne Convention²²⁸ requires member countries to remove registration,

224. See Music Marketplace Study, *supra* note 34, at 93.

225. Mechanical and Digital Phonorecord Delivery Rate Determination Proceeding, 74 Fed. Reg. 4509, 4513 (Jan. 26, 2009) (citation omitted):

The complexity of compliance, and the associated transactions costs, create a curious anomaly: virtually no one uses section 115 to license reproductions of musical works, yet the parties in this proceeding are willing to expend considerable time and expense to litigate its royalty rates and terms. The Judges are, therefore, seemingly tasked with setting rates and terms for a useless license. The testimony in this proceeding makes clear, however, that despite its disuse, the section 115 license exerts a ghost-in-the-attic like effect on all those who live below it. Thus, the rates and terms that we set today will have considerable impact on the private agreements that enable copyright users to clear the rights for reproduction and distribution of musical works.

226. See 2006 Orphan Works Study, *supra* note 37, at 16.

227. See, e.g., Sonny Bono Copyright Term Extension Act of 1998, Pub. L. No. 105-298, 112 Stat. 2827 (codified as amended in 17 U.S.C. §§ 301–304); Council Directive of 29 October 1993 on Harmonizing the Terms of Protection of Copyright and Certain Related Rights, 93/98/EEC, 1993 O.J. (L 290) 9.

228. Berne Convention, *supra* note 30, Article 5.

notice, and any other formalities as preconditions for the enjoyment and exercise of copyright.²²⁹ This requirement practically decreases the incentives for authors to provide sufficient ownership information. Third, digital technology has made it possible to copy and distribute a vast volume of copyrighted works, which quickly multiplies transaction costs required to locate relevant copyright owners.²³⁰

Recent studies demonstrated that orphan work problems are prevalent and have significant effects on public interests.²³¹ For instance, a 2012 survey conducted in the United Kingdom estimated staggering percentages of orphan works:²³² (i) National History Museum, London – 25% of its 500,000 item collection; (ii) European Film Archives – 4% to 7% of its 3.2 million titles; (iii) Imperial War Archive – 20% of its 11 million item collection; (iv) National History Museum, London – 20% of its one million book collection; and (v) National Library of Scotland – around 20% of its 1.5 million book collection. Similarly, several U.S. projects discovered significant shares of orphan works among various library collections, usually in the range of 25% to 50%.²³³

229. See, e.g., Copyright Act of 1976, Pub. L. No. 94-553, § 408(a), 90 Stat. 2541, 2580 (codified as amended at 17 U.S.C. § 408(a)) (registration optional); Berne Convention Implementation Act of 1988, Pub. L. No. 100-568, § 7(a)–(b), 102 Stat. 2853, 2857-58 (codified at 17 U.S.C. §§ 401(a), 402(a)) (notice optional); Copyright Amendments Act of 1992, Pub. L. No. 102-307, § 102(a), 106 Stat. 264, 264 (codified as amended at 17 U.S.C. § 304(a)) (renewal automatic).

230. For cases involving millions of works distributed online, see, for example, *Kelly v. Arriba Soft Corp.*, 336 F.3d 811 (9th Cir. 2003); *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014); *Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015).

231. See, e.g., JISC, *In from the Cold: An Assessment of the Scope of “Orphan Works” and Its Impact on the Delivery of Services to the Public* (Apr. 2009), <http://webarchive.nationalarchives.gov.uk/20140702233839/http://www.jisc.ac.uk/media/documents/publications/infromthecoldv1.pdf> (inter-organizational report on orphan works in the United Kingdom); Anna Vuopala, *Assessment of the Orphan Works Issue and Costs for Rights Clearance* (May 2010), http://www.ace-film.eu/wp-content/uploads/2010/09/Copyright_anna_report-1.pdf (report to the European Commission on the Digital Libraries initiative).

232. See Department of Business, Innovations and Skills (U.K. Intellectual Property Office), *Impact Assessment Report on Orphan Works* (June 2012), at 11, <http://webarchive.nationalarchives.gov.uk/20140603093549/http://www.ipo.gov.uk/consult-ia-bis1063-20120702.pdf>.

233. See 2006 Orphan Works Study, *supra* note 37, at 36–39 (over 50% in selected university libraries); Cairns, *supra* note 42 (25% in the Google Books Project); John P. Wilkin, *Bibliographic Indeterminacy and the Scale of Problems and Opportunities of “Rights” in Digital Collection Building*, RUMINATIONS (Feb. 2011), <http://www.clir.org/pubs/ruminations/01wilkin/wilkin.html> (50% in the HathiTrust corpus).

Many countries in the world have introduced creative legal mechanisms to resolve orphan work problems.²³⁴ This Part is focused on two proposals that may be loosely called the “limitation on liability” approach and the “compulsory license” approach. Additionally, this Part will briefly evaluate various proposals regarding the related but slightly different issue of mass digitization using the Google Books Project as an example.²³⁵

A. LIMITATION ON LIABILITY VERSUS COMPULSORY LICENSE

The U.S. Copyright Office initially deliberated on the “limitation on liability” approach in its well-known 2006 Report on Orphan Works.²³⁶ The report drew inspiration from a 2003 book suggesting that copyright law establish a mechanism similar to marketable title under property law to provide incentives for authors to regularly update ownership information and facilitate copyright transactions.²³⁷ Following these recommendations, Congress introduced multiple bills for orphan works in 2006 and 2008 (none of which have passed).²³⁸ This approach includes the following elements:

(i) A copyrighted work is considered an orphan work where the user is unable to identify and locate the copyright owner with a good faith diligent search;

(ii) A user can start to use the orphan work after filing a notice of use with a government authority;

(iii) If the copyright owner emerges and alleges copyright liabilities, retrospective monetary relief would be limited to reasonable compensation,

234. See, e.g., Axhamn & Guibault, *supra* note 33; Pamela Samuelson, *Legislative Alternatives to the Google Book Settlement*, 34 COLUM. J.L. & ARTS 697, 705 (2011); Commission of the European Communities, *Creative Content in a European Digital Single Market: Challenges for the Future* (Oct. 2009), http://ec.europa.eu/internal_market/consultations/docs/2009/content_online/reflection_paper%20web_en.pdf, at 4.

235. See *Authors Guild v. Google, Inc.*, 804 F.3d 202, 213 (2d Cir. 2015).

236. See 2006 Orphan Works Study, *supra* note 37, at 16. Those recommendations are restated in 2015 Orphan Works Study, *supra* note 21, at 3. The “limitation on liability” approach is different from the argument that the orphan status of a copyrighted work automatically weighs in favor of fair use. The moment that the user is sued for copyright infringement, the works ceases to be an orphan. Therefore, a denial of all remedies, particularly injunctive relief and future royalties, are not justified in many cases.

237. See GOLDSTEIN, *supra* note 79, at 203 (confirmed by a drafter of the report who worked at the U.S. Copyright Office at the time).

238. Shawn Bentley Orphan Works Act of 2008, S. 2913, 110th Cong. (2008); Orphan Works Act of 2008, H.R. 5889, 110th Cong. (2008); Orphan Works Act of 2006, H.R. 5439, 109th Cong. (2006).

i.e., the amount that a willing buyer and a willing seller would have agreed upon before the use began;

(iv) Eligible nonprofit institutions including certain schools, museums, libraries, archives, and public broadcasters would not be subject to any monetary relief for noncommercial uses of orphan works unless they do not promptly cease their uses after receiving notice from the copyright owner; and

(v) A court has the authority to impose injunctive relief upon request of the copyright owner to enjoin further uses of the orphan work accounting for the reliance interests of, and the original expressions added by, the user.

One may argue that the U.S. Copyright Act already contains a watered-down version of the “limitation on liability” approach in the context of mechanical licenses where, if the copyright owner is not locatable in public records, the user may file a notice of intention with the U.S. Copyright Office and then use the work royalty-free until the owner emerges.²³⁹ In 2012, the European Union issued the Directive on Certain Permitted Uses of Orphan Works, which builds on another watered-down version of the U.S. model by providing an exemption for public interest missions of “libraries, educational establishments and museums . . . archives, film or audio heritage institutions and public-service broadcasting organizations.”²⁴⁰

By contrast, Canada²⁴¹ is a celebrated example that has implemented the “compulsory license” approach to tackle orphan work problems, joined by a handful of other countries including the United Kingdom,²⁴² Hungary,²⁴³

239. 17 U.S.C. § 115(b)–(c) (2012).

240. Directive 2012/28 of the European Parliament and of the Council of 25 October 2012 on Certain Permitted Uses of Orphan Works, art. 1(1), 2012 O.J. (L 299) 8 (EU) [hereinafter the E.U. Orphan Works Directive]; see also Australian L. Reform Comm’n, *Copyright and the Digital Economy* (Discussion Paper 79, June 5, 2013), Proposals 12–1, 12–2, 12–3, 12.60, http://www.alrc.gov.au/sites/default/files/pdfs/publications/12_orphan_works.pdf.

241. Copyright Act, R.S.C. 1985, c. C-42, s. 77.

242. Copyright and Rights in Performances (Licensing of Orphan Works) Regulations 2014, S.I. 2014/2863 (U.K.).

243. See Mihály Ficsor, *How to Deal with Orphan Works in the Digital World? An Introduction to the New Hungarian Legislation on Orphan Works*, EUROPEAN PARLIAMENT DIRECTORATE GENERAL FOR INTERNAL POLICIES (2009), <http://www.europarl.europa.eu/document/activities/cont/200911/20091113ATT64497/20091113ATT64497EN.pdf>.

India,²⁴⁴ Japan,²⁴⁵ and Korea.²⁴⁶ Although slight variations exist among these countries, their approach may be summarized based loosely on the Canadian model:

(i) A published work is considered an orphan work where the user is unable to identify and locate the copyright owner with a reasonable effort;

(ii) A user may apply to the Copyright Board (or counterpart authorities in other countries) for a compulsory license to use the orphan work;

(iii) The Copyright Board would grant the compulsory license, usually nonexclusive and nontransferable in nature, if the user sufficiently proves her reasonable effort in searching for the copyright owner;

(iv) The Copyright Board is responsible for establishing the royalty rates and usage conditions for the compulsory license;

(v) The user shall pay her royalties to a collecting society designated by the Copyright Board (or directly to a government agency in several other countries); and

(vi) The collecting society will retain the royalties in escrow while trying to locate the copyright owner for a certain period of time. If the copyright owner does not emerge after the period, the collecting society may allocate the royalties for other purposes, e.g., defraying its administrative costs or funding social and cultural activities.

B. THE INCENTIVE ANALYSIS

The 2012 Chinese Copyright Reform Bill expectedly follows the “compulsory license” approach largely because it would result in rent-seeking opportunities for interest groups like collecting societies:

If the user is unable to identify and locate, with a diligent search, the copyright owner of a published work which copyright has not expired, the user may use the work in a digitalized form after filing an application with, and deposit a royalty at, the organization designated by the copyright administrative authority of the State Council.²⁴⁷

244. The Copyright (Amendment) Act, No. 27 of 2012, Gazette of India (June 7, 2012), http://copyright.gov.in/Documents/CRACKT_AMNDMNT_2012.pdf.

245. Copyright Act, Law No. 48 of 1970, as amended up to Law No. 35 of 2014, art. 67, http://www.cric.or.jp/english/clj/doc/20150227_October,2014_Copyright_Law_of_Japan.pdf (Japan) (unofficial translation).

246. Copyright Act of 1957, Act No. 432, Jan. 28, 1957, as amended up to Act No. 12137, Dec. 30, 2013, art. 50(1), *translated at* http://elaw.klri.re.kr/eng_service/lawView.do?hseq=32626&lang=ENG (S. Kor.) (unofficial translation).

247. Article 51, Third Draft, *supra* note 4.

Notably, evidence from countries already implementing this approach demonstrates that prospective users of orphan works rarely utilize such compulsory licenses. For example, for the twenty-seven-year period from 1988 to 2015, the Copyright Board of Canada has issued only 283 compulsory licenses, which amounts to approximately ten licenses per year.²⁴⁸ This number is drastically dwarfed by the millions of existing orphan works.²⁴⁹

It is easy to understand why the “compulsory license” approach is not popular among potential users, especially compared to the “limitation on liability” approach. First, each time a user wishes to obtain a compulsory license to use an orphan work, she must prove to the satisfaction of the government authority that the copyright owner is not locatable with a good faith diligent search.²⁵⁰ By contrast, with the “limitation on liability,” a user would have to establish her reasonable search in negotiation or in court only if the copyright owner emerges.²⁵¹ Therefore, the “limitation on liability” approach has the advantage of saving administrative costs for both users and adjudicators.

Second, the “compulsory license” approach requires orphan works users to pay copyright royalties upfront, while the “limitation on liability” approach only requires users to pay royalties if the copyright owner resurfaces. Also, it is difficult to imagine how the government authority could efficiently set royalty rates for the compulsory licenses to use orphan works. While governmental rate-setting is by no means easy for any compulsory license, orphan works would add an additional layer of complexity and uncertainty. The government usually strives to simulate the ordinary market value that a willing buyer and a willing seller would have agreed upon through arms-length negotiation.²⁵² Orphan works are by definition out-of-commerce works that do not have existing market benchmarks. The royalty rates for non-orphan works under normal

248. See Copyright Board of Canada, *Decisions – Unlocatable Copyright Owners*, <http://www.cbcda.gc.ca/unlocatable-introuvables/licences-e.html>. The situations appear similar in other countries following the “compulsory license” approach. See, e.g., Marcella Favale et al., *Copyright, and the Regulation of Orphan Works: A Comparative Review of Seven Jurisdictions and a Rights Clearance Simulation* (U.K. Intellectual Property Office, July 2013), <https://www.gov.uk/government/publications/copyright-and-the-regulation-of-orphan-works>, at 46 (eight-two compulsory licenses were granted from 1972 to 2010 in Japan); KOREAN COPYRIGHT OFFICE, <https://www.findcopyright.or.kr/statBord/statBo03List.do?bordCd=3> (ten compulsory licenses were granted by 2015).

249. See *supra* note 232 and accompanying text.

250. See *supra* note 241 and accompanying text.

251. See *supra* note 238 and accompanying text.

252. See, e.g., 17 U.S.C. §§ 112(e)(4), 114(f)(2)(B), 801(b)(1) (2012).

commercial exploitation may not serve as proxy for orphan works, which typically have less market value.²⁵³ Many works become orphaned because copyright owners believe the costs for keeping the ownership information accessible and updated outweigh the benefits of exploiting the works.²⁵⁴

Third, a collecting society designated to manage the compulsory licenses may have enough incentives to collect royalties for orphan works but not enough to locate their copyright owners and distribute royalties. On the one hand, if the money collected is negligible and hardly covers searching costs, the collecting society would naturally be unwilling to search for copyright owners. On the other hand, if the royalties are substantial, the collecting society would have even less incentive to locate copyright owners because it may use the unallocated amount to defray administrative costs and support collective-purpose projects targeting existing members, including awards and scholarships.²⁵⁵

From an efficiency perspective, the “limitation on liability” approach is clearly preferable to the “compulsory license” approach. Copyright protection reflects a trade-off between incentive and access in accordance with the economic features of its subject matters.²⁵⁶ Information products, including works of authorship, have certain characteristics of a public good, i.e., “non-excludability” (or “inappropriability”) and “non-rivalry” (or “indivisibility”).²⁵⁷ “Non-excludability” means that, once information is

253. See Randal C. Picker, *Private Digital Libraries and Orphan Works*, 27 BERKELEY TECH. L.J. 1259, 1283 (2012). To design a compulsory license that encourages users to apply, the government would need to at least ascertain the market values of orphan work licenses (i.e., reasonable damages if liable for infringement) and the probability of copyright owners eventually emerging. It may then set the royalty rates by multiplying the two factors. Above the level, prospective users may choose to proceed without a license.

254. Any financial return from orphan works hardly affects ex ante incentives for creation no matter how successful authors are, especially where the criteria for a diligent search are set sufficiently high and authors have the option to terminate the orphan status. If an author expects her work to be commercially valuable (say 10), she may also expect the probability of her work becoming orphan to be rather low (say 10%). Therefore, the expected value from orphan work licensing would be low (1). If an author otherwise expects the probability of her work becoming orphan is rather high (say 100%), she would rationally anticipate that her work has limited commercial value (1). The expected value is equally low (1).

255. Certain institutional designs may be useful to alleviate the incentive concern; for example, requiring the collecting society to distribute royalties to relevant copyright owners before allocating a percentage of the distributed royalties to defray administrative costs.

256. See *supra* note 41 and accompanying text.

257. See, e.g., ROBERT COOTER & THOMAS ULEN, *LAW AND ECONOMICS* 135 (1988); PAUL A. SAMUELSON & WILLIAM D. NORDHAUS, *ECONOMICS* 37 (17th ed. 2001); William

created and distributed, it is physically difficult to exclude others from enjoying it. The consumption of information is “non-rivalrous” where it may be enjoyed simultaneously by an infinite number of people without incidentally affecting the enjoyment by others. In economic terms, the marginal cost of extending the consumption to another person is near zero. Under such circumstances, it is extremely difficult for authors to recoup the fixed costs of creating their works in a market without property rights because competitors, who are free to copy the same works without incurring the fixed costs, will soon drive the prices towards the marginal costs of reproduction and distribution.²⁵⁸ Therefore, the market tends to undersupply those valuable works absent sufficient incentive for intellectual creation. Copyright law is intended to solve the incentive problem by granting authors exclusive control, for a limited period of time, over the reproduction and distribution of their works, which in turn generates market opportunities for pricing their works above marginal costs. The markup allows authors to recoup their initial investment in creative works, although the increased price may inhibit access by certain consumers who are willing to pay for the marginal cost but not for the premium.

However, orphan works by definition involve authors who may not be located with a diligent search. Chances are that the authors will never reappear. If compulsory licenses are imposed in these cases, users would pay a higher price, but the real authors would not receive any financial incentives. This situation would be the worst of both worlds: limited access for consumers *and* no incentive for authors. By contrast, under the “limitation on liability” approach, consumers would enjoy virtually free access to orphan works, unless the authors reappear and copyright incentives resume functioning properly to signal the authors how much consumers value their works.²⁵⁹ By conditioning certain monetary remedies

M. Landers & Richard A. Posner, *An Economic Analysis of Copyright Law*, 28 J. LEGAL STUD. 325, 326 (1989).

258. From an ex post perspective, once a work is created, the author would be unable to internalize the fixed costs and therefore suffer a competitive disadvantage over free riders who do not bear the fixed costs. From an ex ante perspective, even if the author tries to negotiate a price with all potential users before the work is created, game theory suggests that many users may underbid the work attempting to free ride other consumers' contribution.

259. See GOLDSTEIN, *supra* note 79, at 200 (“[T]here is no better way for the public to indicate what they want than through the price they are willing to pay in the marketplace.”); Harold Demsetz, *Information and Efficiency: Another Viewpoint*, 12 J.L. & ECON. 1, 1 (1969) (arguing that production and consumption of information cannot be judged independently); ADAM SMITH, LECTURES ON JURISPRUDENCE 82–83 (R.L. Meek et al. eds., 1987) (1762) (“[Copyright] is perhaps as well adapted to the real value of the work as any

(e.g., statutory damages and attorneys' fees) on the accessibility of relevant copyright owners, the approach would generate powerful incentives for authors to provide updated ownership information and connect directly with consumers. A well-functioning market would eventually benefit the whole society including consumers and authors alike.

C. MASS DIGITIZATION

The above discussions also suggest that compulsory license and ECL proposals may not be appropriate even for mass digitization projects that involve an astronomical number of copyrighted works and substantial transaction costs for copyright clearance.

Taking the Google Books Project as an example, since 2004 Google has scanned millions of books provided by publishers and libraries.²⁶⁰ On the one hand, Google used the digital corpus to develop book search engines and data mining tools ("non-display uses"). On the other hand, it engaged in negotiation with major publishers and the Authors Guild to launch an online bookstore comparable to Amazon ("display uses"), which Google appeared to discontinue after Judge Chin had rejected the Google Books Settlement.²⁶¹

The Google Books Project has tested the boundaries of copyright protection across the world, especially with regard to non-display uses. In the United States, the Second Circuit recently affirmed the district-court decision that the copying involved in Google's searching and data-mining functions was "transformative use," did not offer the public a meaningful substitute for purchasing copyrighted works, and hence satisfied the test for fair use.²⁶² However, in 2012, a Chinese court found Google liable for full-text book scanning but not for displaying limited snippets.²⁶³ Although

other, for if the book be a valuable one the demand for it in that time will probably be a considerable addition to his fortune. But if it is of no value the advantage he can reap from it will be very small.")

260. See *Authors Guild v. Google Inc.*, 954 F. Supp. 2d 282, 286 (S.D.N.Y. 2013).

261. See *Authors Guild v. Google Inc.*, 770 F. Supp. 2d 666, 679 (S.D.N.Y. 2011).

262. Cf. *Authors Guild v. Google, Inc.*, 804 F.3d 202, 230 (2d Cir. 2015); *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 105 (2d Cir. 2014). While a comprehensive evaluation of Second Circuit decisions regarding the Google Books Project entails a separate article (if not more), it suffices to say at the moment *Authors Guild v. Google* could theoretically arrive at a different outcome than *Authors Guild v. HathiTrust*. In the latter case, the library defendants have originally acquired the copies of relevant books through purchase or other legitimate means. In the former case, Google did not own any copies until it generated new copies through digital scanning. In other words, scanning books eliminates the need for Google but not for libraries to purchase books.

263. See *Wang Xin v. Google, Inc.* (Beijing Interm. People's Ct. Dec. 20, 2012), <http://bjgy.chinacourt.org/paper/detail/2013/10/id/1230596.shtml>.

Chinese law does not contain a four-factor fair use test, the court incorporated part of the three-step test into the reasoning.²⁶⁴ It concluded that the book scanning conflicts with the normal exploitation of copyrighted works because granting licenses and collecting royalties for full-text reproduction is one of the most common exploitations by copyright owners. In addition, the court held that the book scanning unreasonably prejudiced the legitimate interests of copyright owners by creating a potential danger to the market of copyrighted books. Once Google is in possession of the scanned books without authorization, copyright owners would lose control over any subsequent uses of the copies by Google and, if the security system were compromised, the copies would become the breeding bed of countless infringing uses by third parties.²⁶⁵ The Chinese decision appears to attach more importance than the U.S. decision to the market-formation purpose of copyright protection.²⁶⁶

Notably, the clearance difficulty in the Google Books Projects does not result mainly from orphan work issues where searching costs for locating relevant copyright owners are prohibitively high. It has been estimated that merely a fourth of the whole corpus consists of potentially orphan works.²⁶⁷ Google, the largest search engine in the world, had no problem identifying the vast majority of the copyright owners and was actually in the process of negotiating possible licensing agreements with publishers even before the litigation commenced.²⁶⁸ Neither does the sheer volume of copyrighted works involved in the Google Books Project by itself justify a statutory exemption. The increase in transaction costs has been approximately proportionate to the increased volume and increased value of the overall database. It makes little sense to categorically argue that the more copyrighted works a database contains, the less reasonable it is to request a copyright license.

The key barrier to mass digitization appears to be that the incremental value of any individual work to the whole project is often lower than the

264. *See id.* The decision curiously omitted the first step of the three-step test, “limited to certain special cases.”

265. *See id.* The Second Circuit explicitly rejected the hacking scenario as speculative due to “impressive security measures” implemented by Google. *Authors Guild*, 804 F.3d at 228.

266. For the purposes of following sections, mass digitization refers to projects that involve both non-display and display uses. In other words, the discussions are premised on the Google Books Project envisioned in the proposed settlement rather than what Google has implemented so far. Fair use defenses are apparently less plausible for making copyrighted works available online. *See* 2015 Orphan Works Study, *supra* note 21, at 76.

267. *See supra* note 42 and accompanying text.

268. *See* *Authors Guild v. Google Inc.*, 954 F. Supp. 2d 282, 286–87 (S.D.N.Y. 2013).

transaction cost needed to obtain a license for the work.²⁶⁹ For instance, assume that locating a copyright owner takes one dollar, a perfectly reasonable searching cost; the user would still not reach out for a license if scanning the book only added three cents to the whole project.²⁷⁰ This issue is not really different in nature from that faced every day by television/radio broadcasters who use a large number of musical compositions for their programs. If history is any indication, the best solution is not to bypass copyright transactions. Instead, we may pool various copyrighted works together through major publishers or collecting societies to facilitate the issuance of blanket licenses for mass digitization. This approach takes advantage of economies of scale to decrease transaction costs as the formations of ASCAP, BMI, and SESAC do for music rights clearance.²⁷¹

The ECL proposal warrants cautious evaluation as a possible solution to mass digitization for three reasons in particular. First, as discussed above, complicated institutional design would be needed to prevent the ECL from superseding the pricing powers of uninformed nonmember authors, to contain the de jure monopoly of the collecting society against both authors and users, and to create incentives for the collecting society to improve distributional efficiency instead of sitting on royalties collected for

269. Therefore, the Google Books Project actually includes four categories of works: (i) public domain works; (ii) works whose owners opt in; (iii) works whose owners are searchable but searching costs exceed their marginal values; and (iv) orphan works whose owners are not locatable with a diligent search. Apparently, if we define a diligent search by using, as a benchmark, the marginal value of the captioned book, the third and fourth categories would merge into one. *See*, Lang, *supra* note 28, at 135–36.

270. Website search engines provide a familiar example. Although a comprehensive collection of web content is likely the foundation of a search engine, any individual webpage is simply a small portion that may easily be replaced or omitted without any meaningful impact to the overall function of the search engine. Because the search engine actually assists consumers in locating a website, in a competitive market, the website author may be willing to grant a license for free and even pay the search engine to include her website in its search results. Under these circumstances, the license fee is effectively zero or negative. Any search costs, if positive, could appear excessive to the search engine. It may not be efficient to establish a collecting society with the monopolistic power to charge positive prices that create a deadweight loss and substantial administrative costs for handling copyright royalties.

271. In the limited cases where transition costs remain insurmountable and hamper digitization projects, a court may apply a limitation on liability, which however would become unavailable the moment new mechanisms emerge to diminish the transaction costs. It would serve better than a compulsory license to unlock socially valuable utilization of existing works and incentivize future innovations in lowering transaction costs. *Compare* *Am. Geophysical Union v. Texaco, Inc.*, 60 F.3d 913, 917 (2d Cir. 1994) (denying fair use because of new mechanism to lower transaction costs), *with* *Williams & Wilkins Co. v. United States*, 487 F.2d 1345, 1361 (Ct. Cl. 1973) (finding fair use partially because of high transaction costs).

nonmembers.²⁷² The institutional design and its implementation entail substantial costs.²⁷³ Second, the history of the music market demonstrates that competitive collecting societies offering blanket licenses are capable of giving users practically unlimited freedom to use an exhaustive repertoire of relevant works.²⁷⁴ Third, it is unclear why it is even necessary for a project like Google Books to include all of the books in the world to become a viable business. Theoretically, the holdup problem may arise where a project involves a large number of copyright owners and every permission is essential for the whole project to function.²⁷⁵ Under these circumstances, a copyright owner could strategically withhold her permission to increase her share of copyright royalties, which could potentially cause a negotiation breakdown. However, this is usually not the case for mass digitization projects.

Assume that the Google Books Project obtains blanket licenses from collecting societies but accidentally includes a book owned by a nonmember author. If the author claims copyright infringement, Google may remove the infringing work from the digital database and continue its operation with other licensed works (taking comfort in the indemnity provided by collecting societies).²⁷⁶ A single party can hardly have any veto power to block the entire project, which renders the holdup problem remote.²⁷⁷ As a matter of fact, Google has slowed down scanning books from libraries, almost to a complete halt, even though a federal court held that the existing project is exempt from copyright liability as fair use.²⁷⁸ Apparently, the marginal benefit of scanning more books for the purposes of designing book search engines and training web search algorithms quickly diminishes after having scanned 30 million books, although Google

272. See *supra* note 245 and accompanying text. See also David R. Hansen, Kathryn Hashimoto, Gwen Hinze, Pamela Samuelson, & Jennifer M. Urban, *Solving the Orphan Works Problem for the United States*, 37 COLUM. J. L. & ARTS 1, 46 (2013).

273. For example, the U.K. Intellectual Property Office has estimated the cost of establishing the supervising authority would be £2.5 million to £10 million and the costs of operating the supervising authority would be £0.5 million to £1.8 million annually. See Intellectual Property Office, *supra* note 232, at 6.

274. See Music Marketplace Study, *supra* note 34, at 19.

275. See *supra* note 44 and accompanying text.

276. See *supra* note 110 and accompanying text.

277. For recent articles that discuss the holdup problem, see John M. Golden, Commentary, “Patent Trolls” and Patent Remedies, 85 TEX. L. REV. 2111, 2139 (2007); Mark A. Lemley & Carl Shapiro, *Patent Holdup and Royalty Stacking*, 85 TEX. L. REV. 1991, 1993 (2007); J. Gregory Sidak, *Holdup, Royalty Stacking, and the Presumption of Injunctive Relief for Patent Infringement: A Reply to Lemley and Shapiro*, 92 MINN. L. REV. 714, 716 (2008).

278. See Howard, *supra* note 45.

announced in 2010 that there are a total of 130 million books in the world (129,864,880 to be precise).²⁷⁹

The ECL proposal for mass digitization is sometimes justified on the premises that an opt-out choice is available for nonmember authors, and transaction costs would be lower for *some* authors to opt out of a project than for a prospective user to approach *all* relevant authors.²⁸⁰ It may appear to be no more than a small inconvenience for an author to opt out of the Google Books Project by approaching the proposed Book Rights Registry, a single collective society intended to manage a single project.²⁸¹ However, if the ECL system becomes widespread in numerous countries and mass digitization projects proliferate for various purposes, an author striving to exploit her exclusive rights worldwide would have to monitor multiple projects managed by multiple collecting societies around the world. If she wishes to opt out of the ECL regimes and instead manage some or all her works by herself, she must carefully comply with opt-out requirements set by different countries. These daunting tasks are exactly the kind of formalities that the drafters of the Berne Convention envisioned while determining to completely prohibit any formality as a precondition for the enjoyment and exercise of exclusive rights.²⁸²

V. CONCLUSION

Digital technologies have significantly lowered the costs involved in producing, marketing, and distributing copyrighted content. Online service providers such as Amazon, YouTube, and Spotify have the necessary economic and technological capacities to make an enormous volume of multimedia content available to the general public. However, due to relatively limited innovation in the field of rights clearance,²⁸³ these online service providers are still facing substantial transaction costs in tracking

279. See Taycher, *supra* note 45.

280. See 2015 Orphan Works Study, *supra* note 21, at 93.

281. See, e.g., Daniel Gervais, *The Changing Role of Copyright Collectives, in COLLECTIVE MANAGEMENT OF COPYRIGHT AND RELATED RIGHTS 26* (Daniel Gervais ed., 2006) (“If it is a restriction at all, [ECL] is a mild one. It guarantees an orderly exploitation of the repertoire that will be licensed but offers authors the option of going back to Level 0 by sending a simple notice, perhaps even as simple as an email.”).

282. Berne Convention, *supra* note 30, Article 13.

283. Notably, several projects have recently emerged to streamline copyright clearance in the digital environment. See, e.g., *Copyright Policy Practicum*, <https://law.stanford.edu/education/only-at-sls/law-policy-lab/practicums-2014-2015/copyright-policy-practicum/>; *Copyright Hub*, <http://www.copyrighthub.co.uk/>; *Google Content ID*, <https://support.google.com/youtube/answer/2797370?hl=en>.

down copyright owners, negotiating license terms, and acquiring proper authorizations for millions of different copyright works.²⁸⁴ Policymakers around the world are working to unlock copyrighted works in the digital age. Copyright reform initiatives largely follow one of two directions: Policymakers may change copyright from an opt-in regime into an opt-out or all-in regime (e.g., ECLs and compulsory licenses) to eliminate the necessity of copyright transactions and allow downstream users to exploit copyrighted works without authorization. Alternatively, policymakers may streamline private transactions in the marketplace, create incentives for authors to provide licensing information, and eventually allow market players to innovate on efficient business models. In comparison to the market approach, compulsory licenses have a number of drawbacks, such as divesting authors of exclusive rights in copyrighted works, resulting in wasteful rent seeking and setting arbitrary prices for copyright royalties. However, the fundamental concern is that compulsory licenses would undermine the incentives for collecting societies and other market players to improve their services in order to decrease transaction costs. While the United States and the rest of the world are at a crossroads in copyright reform, the road taken (and the road not taken) by Chinese policymakers provides a valuable lesson: We cannot, in the name of lowering transaction costs, completely sidestep transactions and sidestep the market as the principal mechanism to allocate social resources for intellectual creation.

284. See, e.g., James Duffett-Smith, *Comments OF Spotify USA Inc.*, http://copyright.gov/docs/musiclicensingstudy/comments/Docket2014_3/Spotify_USA_1nc_MLS_2014.pdf. (Spotify spent almost three years for clearance in the United States.).