

JOURNAL OF LAW, ECONOMICS & POLICY

VOLUME 9

FALL 2012

NUMBER 1

EDITORIAL BOARD 2012-2013

Adam Park
Editor-in-Chief

Nathaniel Harris
Executive Editor

Morgan Shields
Publications Editor

Benjamin Charlton
Managing Editor

Genevieve Miller
Senior Articles Editor

Carolyn Head
Senior Notes Editor

Jennifer Johnston
Senior Research Editor

Peter Anderson
Kylie Caporuscio
Anthony Kanakis
Articles Editors

Emily Barber
Juan Kassar
Mark Stevens
Notes Editors

David Brodian
Jason Malashevich
Tom Randall
Research Editors

Anthony Kanakis
Articles Submissions Editor

Tom Randall
Technology Editor

Mark Weiss
*Associate Articles
Submissions Editor*

Jacob Merrill
Outreach Editor

MEMBERS

Michael Brody
Hollie Kapos

Louise Martin
Elizabeth Newell
Adam Schneider

Robert Strange
Lauren Wynns

CANDIDATE MEMBERS

Josh Branson
Marie Keiko Breyan
Victoria Clarke
Brady Cummins
David Dubin
Steven Dunn
Meagan Dziura
Kelly Fryberger

Kevin Hill
Ryan Leavitt
Sarah Mernin
Mariam Noori
Catherine Oryl
Erica Peterson
Michael Rogers

Ian Rothfuss
Maxwell Slackman
Kyra Smerkanich
Jonathan Stapor
Charles Thresher
Allison Walsh
Preston Wise
Crystal Yi

JOURNAL OF LAW, ECONOMICS & POLICY

VOLUME 9

FALL 2012

NUMBER 1

BOARD OF ADVISORS

Lisa E. Bernstein
James M. Buchanan
Judge Guido Calabresi
Lloyd R. Cohen
Robert D. Cooter
Robert C. Ellickson
Richard A. Epstein
Judge Douglas H. Ginsburg
Mark F. Grady
Bruce H. Kobayashi
Henry G. Manne
A. Douglas Melamed

Francesco Parisi
Eric Posner
Judge Richard A. Posner
Roberta Romano
Hans-Bernd Schäfer
Steven M. Shavell
Henry E. Smith
Vernon L. Smith
Gordon Tullock
Thomas S. Ulen
W. Kip Viscusi
Todd J. Zywicki

JOURNAL OF LAW, ECONOMICS & POLICY

VOLUME 9

FALL 2012

NUMBER 1

CONTENTS

ARTICLES

- 1 DESIGN, INSTITUTIONS, AND THE EVOLUTION OF PLATFORMS
Richard N. Langlois
- 15 PLATFORMS, TEAMWORK AND CREATIVITY: MEDIATING HIERARCHS IN THE NEW ECONOMY
Salil Mehra
- 29 THE BROTHERS GRIMM BOOK OF BUSINESS MODELS: A SURVEY OF LITERATURE AND
DEVELOPMENTS IN PATENT ACQUISITION AND LITIGATION
Anne Layne-Farrar
- 59 THE PRIVATE COSTS OF PATENT LITIGATION
James Bessen & Michael J. Meurer
- 97 NEXT-GENERATION COMPETITION: NEW CONCEPTS FOR UNDERSTANDING HOW INNOVATION
SHAPES COMPETITION AND POLICY IN THE DIGITAL ECONOMY
David J. Teece

COMMENTS

- 119 COLLATERAL CONSEQUENCES OF CRIMINAL CONVICTIONS: A COST-BENEFIT ANALYSIS
Genevieve J. Miller
- 145 EMPOWERING LOCAL AND SUSTAINABLE FOOD: DOES THE FOOD SAFETY MODERNIZATION ACT'S
TESTER-HAGAN AMENDMENT REMOVE ENOUGH BARRIERS?
Peter Anderson

DESIGN, INSTITUTIONS, AND THE EVOLUTION OF PLATFORMS

*Richard N. Langlois**

ABSTRACT

Baldwin and Woodward¹ define a platform as “a set of stable components that supports variety and evolvability in a system by constraining the linkages among the other components.” Musing on this definition, I look at the evolution of platforms from an economic and historical perspective rather than a strictly strategic or product-design one. I point to three interacting and interdependent constellations of forces that shape the development of platforms: the extent of the market, institutions, and strategic design. Although the extent of the market comes closest to being an exogenous factor, it too is shaped—at least in the short run—by institutional and strategic vectors.

In his entertaining—and, to my taste, entirely right-headed—new book *Civilization*, the historian Niall Ferguson² ascribes the economic, political, and cultural domination of the West in the modern period to six “killer apps”: competition, science, property rights, medicine, the consumer society, and work ethic. What Ferguson really means, of course, is that Europe developed systems of economic, political, and cultural institutions that gave it an advantage over other regions. Catchy though it be, the term “app” here is probably inapt. What Ferguson means is that Europe developed a set of institutional *platforms* for which Europeans could write myriad apps: global exploration, the theory of evolution, the enclosure movement, pasteurization, mantel clocks, and factory discipline.³

To many readers, I suppose, calling things like competition and science platforms is just as flamboyantly fuzzy-headed as calling them apps. In most discussions, the concept has a narrower meaning. In economics, for example, a platform is a generalization of the idea of a two-

* Professor of Economics, The University of Connecticut, Storrs, CT 06269-1063 USA, (860) 821-0152 (phone), (860) 486-4463 (fax), Richard.Langlois@UConn.edu, <http://langlois.uconn.edu/>. Keynote address for the conference “The Digital Inventor: How Entrepreneurs Discover, Profit, and Compete on Platforms,” *Journal of Law, Economics & Policy*, Information Economy Project, Friday, February 24, 2012, George Mason University School of Law, Arlington, Virginia.

¹ Carliss Y. Baldwin & C. Jason Woodward, *The Architecture of Platforms: A Unified View*, in PLATFORMS, MARKETS AND INNOVATION 19, 19 (Annabelle Gawer ed., 2009).

² NIALL FERGUSON, CIVILIZATION: THE WEST AND THE REST 12-13 (2011).

³ To be fair, perhaps we could see Ferguson’s six “apps” as running on some higher-level platform, maybe the network of human cognition. By the same token, many of the “apps” running on the six platforms are themselves in turn platforms running apps (just as a browser is an application on a computer platform but is at the same time a platform that runs browser apps). And so on down the line.

sided (or multisided) market.⁴ In the context of strategy and product design, a platform is a modular configuration of technological elements that permits rapid reconfiguration.⁵ These two meanings are more closely related than may at first appear, and they may even be related in the end to the kind of platforms about which Ferguson is writing.

All markets have at least two sides in that markets coordinate the needs of both buyers and sellers. A platform comes into the picture when an entity must simultaneously coordinate buyers and sellers with special problems of complementarity. Consider two alternative methods of coordinating grain farmers and flour millers. One could imagine a wholesaler who buys grain from farmers and sells it to the millers. Alternatively, one could imagine a system in which farmers and millers trade lots of grain directly through an elaborate institutional structure involving standardization and bidding. Economists would not want to call the first system a two-sided market (or a platform): it is just an ordinary market with an intermediary. But the second coordination system—the Chicago Mercantile Exchange, as it turns out—might well qualify as a platform. In accordance with the definition offered by Hagiu and Wright,⁶ a mercantile exchange of this sort is indeed a platform because it facilitates *direct* interaction between *affiliated* players. Or, as Rysman puts it,⁷ a market becomes a platform when two sets of agents interact through an intermediary, and the decisions of each set of agents affect the outcomes of the other set of agents, typically through an externality.⁸

In the American West in the nineteenth century, the coming of the railroad spurred a transformation from something like an ordinary market to something like a platform.⁹ With the railroad came another innovation: the grain elevator. Wheat from many farmers would thus be dumped into the same large hopper, and buyers could no longer assure quality by trusting the reputations of known individual farmers. So, in addition to providing an institutional structure for trade in anonymous lots, the Chicago Mercan-

⁴ See Marc Rysman, *The Economics of Two-Sided Markets*, 23 J. ECON. PERSP. 125, 125 (2009); see also Jean-Charles Rochet & Jean Tirole, *Platform Competition in Two-Sided Markets*, 1 J. EUR. ECON. ASS'N 990, 990 (2003).

⁵ Baldwin and Woodward define a platform as “a set of stable components that supports variety and evolvability in a system by constraining the linkages among the other components.” Baldwin & Woodward, *supra* note 1.

⁶ Andrei Hagiu & Julian Wright, *Multi-Sided Platforms* 9-10 (Harv. Bus. Sch., Working Paper No. 12-024, 2011).

⁷ Rysman, *supra* note 4.

⁸ The extent to which these externalities are technological rather than pecuniary is, however, a matter of controversy. See S.J. Leibowitz & Stephen E. Margolis, *Network Externality: An Uncommon Tragedy*, 8 J. ECON. PERSP. 136-137, 149 (1994). I will talk about network *effects* or “special coordination problems.”

⁹ See WILLIAM CRONIN, *NATURE'S METROPOLIS: CHICAGO AND THE GREAT WEST* 116-19 (1991).

tile Exchange also had to develop technological standards for type and quality of grain to facilitate monitoring. Notice that this example of a platform does have some affinities with the concept of a reconfigurable modular system. It is in essence a technological standard—a standard that defines the category of Hard Red Winter wheat, for example—that permits trading in modular units of 5,000 bushels.

The economics literature on platforms and the strategy-technology literature on platforms share the influence of an earlier literature on technical standards and network effects.¹⁰ Here complementarities and special coordination problems stand out in sharp relief. The would-be typist's choice of which touch-typing system to learn depends crucially on which keyboard layout she expects most employers to choose; at the same time, the employer's choice of keyboard layout depends crucially on which set of touch-typing skills are likely to be in greatest abundance. Or, consider an only-slightly-less-archaic example: the user's choice of computer operating system depends crucially on which system he or she expects to attract the largest number of "apps" (once laboriously referred to as "applications" or "programs"); at the same time, the likelihood that a developer chooses to write apps for a particular operating system depends crucially on how many people he or she expects will adopt the system. The standards that link together typist and keyboard, OS and apps, constitute a platform.¹¹

The present-day economics literature of multisided platforms takes its cue from this earlier literature on standards and path dependency.¹² Despite the importance of (technological) standards in the path-dependency literature, the focus was not on the platform's technological structure but on the nature of the market coordination problem caused by the fact of complementarity and standardization. The present-day economics literature has run with the market part of the story, generalizing that story to problems of coordination under complementarity that do not look very technological or

¹⁰ See, e.g., Paul A. David, *Clio and the Economics of QWERTY*, 75 AM. ECON. REV. 332 (1985); Joseph Farrell & Garth Saloner, *Standardization, Compatibility, and Innovation*, 16 RAND J. ECON. 70 (1985); Michael L. Katz & Carl Shapiro, *Network Externalities, Competition, and Compatibility*, 75 AM. ECON. REV. 424 (1985); S.J. Liebowitz & Stephen E. Margolis, *Path Dependence, Lock-In, and History*, 11 J.L. ECON. & ORG. 205 (1995); S.J. Liebowitz & Stephen E. Margolis, *The Fable of the Keys*, 33 J.L. ECON. 1 (1990).

¹¹ In this literature, the term *standard* is used capaciously to embrace all of what Baldwin and Clark call design rules: the *architecture* of the platform, the *interfaces* that connect its components, and the *standards*, narrowly understood, that test conformance with the architecture and interfaces. See CARLISS Y. BALDWIN & KIM B. CLARK, *DESIGN RULES. VOL. 1: THE POWER OF MODULARITY* 225 (2000).

¹² "Indeed, in a technical sense, the literature on two-sided markets could be seen as a subset of the literature on network effects." Rysman, *supra* note 4, at 127.

even standardized at all. A common example is the problem of equating the number of men and women who mingle at a singles bar.¹³

The focus of attraction in the path-dependency literature was whether such lock-in leads to the perpetuation of an inefficient standard—that is, to a “market failure.”¹⁴ In the context of a proprietary standard, some feared—and some continue to fear—that lock-in might lead to monopoly. Indeed, some have argued that a proprietary standard is a natural-monopoly bottleneck, like a critical railroad bridge or a pipeline, which should fall under the “essential facility” doctrine of antitrust law.¹⁵ In Schumpeter’s famous image, however, competition is a “perennial gale of creative destruction.”¹⁶ And, as David Teece and others have insisted, the rents we observe in this windswept kind of competition are at best quasi-rents—that is, temporary scarcity rents—and are more likely “Schumpeterian” rents—that is, temporary returns to entrepreneurship and innovation.¹⁷

Clearly, the narrower the scope of a technological standard, the more temporary—the more “Schumpeterian”—the rents are likely to be. For example, major personal computer applications like word processors and spreadsheets involve technical standards, and competition among such programs involves network effects. This has led to dominant applications in the various program categories, and the owners of those dominant programs have presumably enjoyed rents during the period of dominance. But those periods have historically been relatively brief, as new dominant programs embodying a new standard came to displace their predecessors in a process of “serial monopoly.”¹⁸ Even when platform standards are relatively wide in scope and seemingly durable, however, it may well be that competition among platforms remains the superior alternative, especially if one refuses to see antitrust and other forms of regulation as disinterested and costless.

Moreover, as the literature on network effects has morphed into the literature of multisided platforms, a clear message for policy has emerged: *it is complicated*. Informed by simple neoclassical models of competition, antitrust policy in the decades after World War II assumed a posture that

¹³ David S. Evans, *The Antitrust Economics of Two-Sided Markets*, 20 YALE J. ON REG. 325, 332-33 (2003).

¹⁴ See Arthur, W. Brian, *Competing Technologies, Increasing Returns, and Lock-in by Historical Events*, 99 ECON. J. 116, 128 (1989); see also David, *supra* note 10, at 332-37.

¹⁵ Richard N. Langlois, *Technological Standards, Innovation, and Essential Facilities: Toward a Schumpeterian Post-Chicago Approach*, in DYNAMIC COMPETITION AND PUBLIC POLICY: TECHNOLOGY, INNOVATION, AND ANTITRUST ISSUES 193, 203 (Jerry Ellig ed., 2001).

¹⁶ JOSEPH A. SCHUMPETER, CAPITALISM, SOCIALISM AND DEMOCRACY 84 (Harper & Bros., 2d 1947).

¹⁷ David J. Teece & Mary Coleman, *The Meaning of Monopoly: Antitrust Analysis in High-Technology Industries*, 43 ANTITRUST BULL. 801, 820-22 (1998).

¹⁸ S.J. LIEBOWITZ & STEPHEN E. MARGOLIS, WINNERS, LOSERS & MICROSOFT: COMPETITION AND ANTITRUST IN HIGH TECHNOLOGY 15 (1999).

Williamson famously branded “the inhospitability tradition.”¹⁹ In the real world, unlike the world of the simple models, competitors engage in a wide variety of strategies, and those strategies often involve restrictions on others and deviations from simple marginal-cost pricing. The inhospitability tradition (a) viewed such strategies as *ipso facto* anticompetitive, and (b) assumed without thought that antitrust policy would be a knife sharp and subtle enough to correct the problem.²⁰ By studying carefully the economics of multisided platforms, the new literature has illuminated the ways in which various seemingly inexplicable strategies are actually aimed at solving complex problems of coordination.²¹ Cross-subsidies (including cross-subsidies over time), tying, price discrimination, and vertical and horizontal restraints are often rational and reasonable strategies in the context of multisided platforms. To the extent that such strategies generate economic rents—as they inevitably will—it is far from clear that allowing competitors free rein is not a better second-best solution than antitrust and regulatory tinkering.²²

From a larger perspective, of course, establishing a multisided platform is itself a strategy. The multisidedness of a market is often not predetermined by technology but is an endogenous choice.²³ It is here that the literature on technology and strategy of modular systems picks up from the economics of multisided platforms.

Researchers interested in strategy and product design—and even the occasional errant economist—have looked at platforms as the solution to what is arguably an even more fundamental problem of coordination: the coordination of complexity. In a complex system, many parts must work together.²⁴ If the interaction among the parts is haphazard, the costs of coordinating the many elements can outweigh the benefits of specialization.²⁵ But by slicing the system in a clever way, one can create a modular system that hides or *encapsulates* complexity.²⁶ With complexity safely behind the impermeable boundaries of modules, coordination can take place

¹⁹ Oliver E. Williamson, *Assessing Vertical Market Restrictions: Antitrust Ramifications of the Transaction Cost Approach*, 127 U. PA. L. REV. 953, 959 (1979).

²⁰ The assumption that antitrust and regulatory policies operate costlessly is the flip side of the assumption that markets can or should operate costlessly.

²¹ See generally Rysman, *supra* note 4, at 125, 129-37.

²² See generally Evans, *supra* note 13, at 325, 380.

²³ See generally Rysman, *supra* note 4, at 125, 132-35.

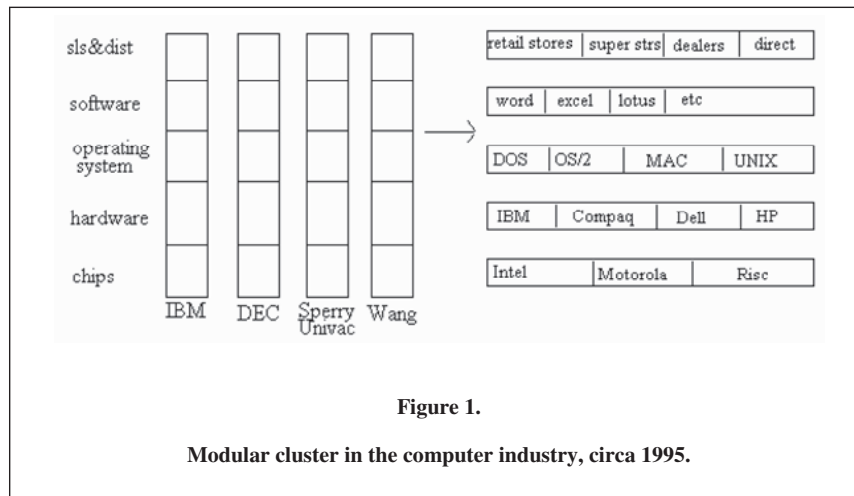
²⁴ See Gary S. Becker & Kevin M. Murphy, *The Division of Labor, Coordination Costs, and Knowledge*, 107 Q. J. ECON. 1137, 1137-38 (1992).

²⁵ *Id.* at 1156-57; FREDERICK P. BROOKS, JR., *THE MYTHICAL MAN-MONTH: ESSAYS ON SOFTWARE ENGINEERING 7* (Addison-Wesley Prof'l, 2d ed. 1995) (1975).

²⁶ See Herbert A. Simon, *The Architecture of Complexity*, 106 PROC. AM. PHIL. SOC'Y, 467, 473-77 (1962) (calling such a system “decomposable”); see also BALDWIN & CLARK, *supra* note 11, at 59; Richard N. Langlois, *Modularity in Technology and Organization*, 49 J. ECON. BEHAV. & ORG. 19, 20-21 (2002).

through lean interfaces, with most of the action happening at thin crossing points, that is, at places of connection that do not require transmitting dense or complex information. The emergence of a platform involves more than the (modular) design of a product. As Simon suggests, platform emergence is about the structure of *decision-making* as well, about the way in which the process production is organized.²⁷

The emergence of a platform can be transformative on a Schumpeterian scale. Andy Grove²⁸ famously pointed out that the modular platform of the personal computer (PC) had led to a fundamental transformation of industry structure, from a world of vertically integrated computer firms—vertical “silos”—to a network of vertically specialized entities—a “modular cluster.”²⁹ See Figure 1.³⁰



The home mortgage industry is another, and perhaps timelier, example of a transformation from silos to a modular cluster. Until the late twentieth century, the various steps involved in residential mortgages—identifying

²⁷ As is often the case in the study of organization, Herbert Simon explained the matter well a long time ago: “Any division of labor among decisional sub-systems creates externalities, which arise out of the interdependencies among the subsystems that are ignored. What is wanted is a factorization that minimizes these externalities and consequently permits a maximum degree of decentralization of final decision to the subsystems, and a maximum use of relatively simple and cheap coordinating devices like the market mechanism to relate each of the decisional subsystems with the others.” Herbert A. Simon, *Applying Information Technology to Organization Design*, 33 PUB. ADMIN. REV. 268, 270 (1973).

²⁸ ANDREW S. GROVE, *ONLY THE PARANOID SURVIVE: HOW TO EXPLOIT THE CRISIS POINTS THAT CHALLENGE EVERY COMPANY AND CAREER* 41-45 (1996).

²⁹ BALDWIN & CLARK, *supra* note 11, at 16.

³⁰ GROVE, *supra* note 28, at 44.

lenders, identifying and sorting borrowers, assessing creditworthiness, closing the loan, servicing the loan, and making payments to the lenders—all took place within integrated retail banks or savings and loan associations.³¹ Beginning in the 1970s, the individual pieces of this industry began to break apart. As the world has since learned, this transformation involved packaging mortgages together to create securities for trade in financial markets. Standards played a critical role—sorting loans into established categories according to the creditworthiness of borrowers (and other attributes) much the way the standards of the Chicago Mercantile Exchange had sorted wheat. In both cases, the standards facilitated high-volume anonymous trade. In both cases, the standards substituted for a more costly system of quality assurance, rooted in personal relations and often-idiosyncratic knowledge. As Jacobides emphasizes, the vertically-specialized mortgage system took advantage of the principle of hiding complexity.

The key to making securitization work was to create a structure in which mortgagors would not know who really owned their loans; they would be dealing with the mortgage banks that originated them. Indeed, many of the readers of this paper may not know that their mortgages are really held by a smattering of financial institutions around the globe. Likewise, owners of the assets do not have any idea of the specific details of the loans they own. More important, the securitizers do not have to coordinate with the mortgage banks on anything more than receiving the payments from the loans at the prespecified intervals; the obligations, in case of foreclosure, are well defined ex ante; there is no need for any communication about the loans among the originator/servicer, securitizer, and owner.³²

The benefits of efficient complexity hiding appear not only in the form of more finely tuned coordination between buyer and seller but also—and importantly—in the form of innovation. A modular architecture encourages modular innovation: improving the internal structure of the complex modules, rearranging and recombining those modules, and adding new modules to the system.³³ Because a modular structure can engage rapid trial-and-error learning, and can take advantage of more potential sources of ideas, such a structure can unleash what Baldwin and Clark call the option value of a modular system.³⁴

From a distant perspective, the high option value of a modular system should make inevitable the tectonic shift from vertical silos to modular

³¹ See Michael G. Jacobides, *Industry Change through Vertical Disintegration: How and Why Markets Emerged in Mortgage Banking*, 48 ACAD. MGMT. J. 465, 469 (2005).

³² *Id.* at 479.

³³ See BALDWIN & CLARK, *supra* note 11, at 92; see also Raghu Garud & Arun Kumaraswamy, *Technological and Organizational Designs for Realizing Economies of Substitution*, 16 STRATEGIC MGMT. J. 93, 94-95, 106 (1995); Richard N. Langlois & Paul L. Robertson, *Networks and Innovation in a Modular System: Lessons from the Microcomputer and Stereo Component Industries*, 21 RES. POL'Y 297, 301-02 (1992).

³⁴ See BALDWIN & CLARK, *supra* note 11, at 92.

clusters. In the long run, as Smith³⁵ insisted, an increase in the extent of the market demands an increase in the division of labor, including the organizational division of labor.³⁶ In any shorter run, however, we are left with the question of whether and when modular transformations will take place. Ultimately, the organizational implications of modularity must conform to the logic of institutional change more generally. Ruttan and Hayami rightly insist that we must look not only at the *demand* for institutional change but also at its *supply*.³⁷ The demand for institutional change comes from the increase in total value an institutional innovation would generate. In the case of a modular cluster, the increase in option value constitutes a pile of \$5 bills lying on the ground. But supplying institutional change—seizing those tantalizing bills—depends on overcoming a variety of costs generated by the pre-existing structure of institutions. It is here, in analyzing these costs, that the economics of multisided markets and theories of strategic platform design come together.

Strategy scholars have long understood that organizational choices—choices about the boundaries of ownership—are strategic variables. Competitors experiment with organizational forms in order to find and appropriate economic rents.³⁸ In this context, for example, vertical integration and a multisided market might be strategic alternatives. Nowadays, the Android and iOS platforms coordinate arms-length transactions between affiliated consumers and app developers. Android also coordinates between affiliated consumers and hardware makers, whereas Apple produces all its own hardware. In the early days of the handheld organizer, Palm created its own hardware, operating system, and apps in-house, selling an even more bundled product to consumers.³⁹ Individual players presumably choose these alternative organizational arrangements in an effort to appropriate the largest rents.

At first glance, one might think that keeping as many functions as possible in-house would lead to the greatest appropriation of rents. Notice, first of all, that there are two (often intertwined) ways of doing this: technological design and property rights. The so-called Wintel standard—known in the dim past as the IBM-compatible standard—in PCs coordinates between affiliated consumers and hardware producers down to the modular

³⁵ See ADAM SMITH, AN INQUIRY INTO THE NATURE AND CAUSES OF THE WEALTH OF NATIONS 8 (Oxford: Clarendon Press 1976) (1776).

³⁶ See Richard N. Langlois, *The Vanishing Hand: The Changing Dynamics of Industrial Capitalism*, 12 INDUS. & CORP. CHANGE 351, 352 (2003).

³⁷ See Vernon W. Ruttan & Yujiro Hayami, *Toward a Theory of Induced Institutional Change*, 20 J. DEV. STUD. 203, 213-15 (1984).

³⁸ As has often been noted, what is a vice in antitrust economics—protected economic rents—becomes a virtue in the context of strategy. See generally David J. Teece, *Profiting from Technological Innovation: Implications for Integration, Collaboration, Licensing, and Public Policy*, 15 RES. POL'Y 285, 291, 304 (1986).

³⁹ See generally Rysman, *supra* note 4, at 125, 132-33.

subsystems level. This came about because IBM lost control of the set of standards that coordinate among the subsystems, and those standards passed into the public domain.⁴⁰ By contrast, Apple produces its own hardware. It does so in part because, having retained ownership of the relevant technological standards, it can refuse to license to others the right to construct subsystems. In addition, however, the architecture of the Macintosh is less modular in crucial ways, making it harder for outsiders to create compatible subsystems.⁴¹ The PC has always relied on a relatively simple Basic Input-Output System (BIOS) to coordinate between the hardware and the operating system, whereas the same functions in the Macintosh are more complex and firmly integrated into the operating system. IBM lost control of its standards largely because the PC BIOS was simple and could be reverse-engineered without violating copyright.⁴²

So is this the message for strategic rent-seekers: make your system non-modular and make sure you keep control of the relevant property rights? The problem is that if Baldwin and Clark are right—and they are—the option value of a modular system is so great that it generates and places on the ground far more \$5 bills in total than a non-modular alternative. This creates strategic pressure along several dimensions. Ultimately, that pressure may easily erupt to destroy the non-modular strategy.

A case in point is the consumer electronics industry in the early twentieth century, especially the radio and related audio reproduction technology.⁴³ For a variety of reasons, including a fear of what we would nowadays call a patent anticommons,⁴⁴ the U.S. federal government created a single national champion in the radio industry: the Radio Corporation of America (RCA). The radio was in fact a relatively modular system—something easy to assemble from a variety of relatively standard parts. In this respect, the early radio resembled the early PC, including the significance in both cases of a large and dispersed hobbyist community. RCA united ownership of the patents for key components like the diode, the triode, and the superheterodyne circuit, which in principle reduced the transaction costs of finding and bribing the patent holders. But as soon as its patents were declared valid by the courts, RCA embarked on a strategy of package licensing.⁴⁵

⁴⁰ Richard N. Langlois, *External Economies and Economic Progress: The Case of the Microcomputer Industry*, 66 BUS. HIST. REV. 1, 29 (1992). That is to say, the PC bus standard is now controlled by a subcommittee of the IEEE, the electrical engineering trade association.

⁴¹ Baldwin & Woodward, *supra* note 1, at 35.

⁴² Langlois, *supra* note 40, at 23-24.

⁴³ Richard N. Langlois, *Organizing the Electronic Century*, in THE THIRD INDUSTRIAL REVOLUTION IN GLOBAL BUSINESS (forthcoming Jan. 2013).

⁴⁴ Michael A. Heller & Rebecca S. Eisenberg, *Can Patents Deter Innovation? The Anticommons in Biomedical Research*, 280 SCI. 698, 698-701 (1998).

⁴⁵ Initially only twenty-five large assemblers would have the rights to RCA's patents, in exchange for a sizeable royalty of 7.5% plus back damages for infringement. Although RCA did later extend the

The licensing was a package in the sense that an assembler had to pay royalties on RCA patents for all relevant parts of the radio, even if the assembler did not use all those parts. As Graham has noted, “the most enduring consequence of the [package licensing] policy was that it made it uneconomic for most other companies to do radio-related research, because they could not recoup the investment. This left control of the rate and direction of technological change in the radio industry largely in the hands of RCA.”⁴⁶

Reversing the judgment of Alfred Chandler—that the integrated structure of RCA and firms like it, along with the related benefits of large research-and-development labs, drove innovation in the early twentieth century⁴⁷—I have suggested instead that RCA’s integrated structure failed ultimately to tap the option value of what was potentially a powerful modular architecture.⁴⁸ By the second half of the twentieth century, RCA’s integrated strategy became vulnerable to focus attacks, especially from Japan, on pieces of the system. Japanese firms concentrated on entry-level elements like transistor radios and black-and-white televisions, parlaying these into wider competition with RCA and other integrated American firms as their capabilities grew. They also focused on pieces of the system, like videotape recording, that American firms had relatively neglected. By the 1980s, the American consumer electronics industry and its large integrated firms were no more.⁴⁹

This analysis of RCA and the American consumer electronics industry is of course a variant of, and was inspired by, the Baldwin and Clark argument about another post-war industry: mainframe computers.⁵⁰ By maintaining a closed standard within a vertical silo, they argue, IBM—the dominant force in the industry—forewent the huge option value of the 360 System of computers.⁵¹

The 360 was very much a modular system—that was its very *raison d’être*. Already a leader in computers in the 1950s, IBM had found the costs mounting alarmingly of coordinating the software necessary for the myriad special-purpose computing machines they produced. By creating a single, highly reconfigurable modular system, the company hoped to slice

deal to others and reduce royalty demands somewhat, it was nonetheless RCA’s control of the patent portfolio that gave shape to the radio industry. Langlois, *supra* note 43.

⁴⁶ MARGARET B.W. GRAHAM, *RCA AND THE VIDEODISK: THE BUSINESS OF RESEARCH* 41 (Cambridge Univ. Press 1986).

⁴⁷ ALFRED D. CHANDLER, *INVENTING THE ELECTRONIC CENTURY: THE EPIC STORY OF THE CONSUMER ELECTRONICS AND COMPUTER INDUSTRIES* (Free Press 2001).

⁴⁸ Langlois, *supra* note 43.

⁴⁹ It also did not help, as Chandler rightly notes, that RCA and its competitors had lost focus because of their dependence on royalty income and the distraction of defense-related research and development. CHANDLER, *supra* note 47.

⁵⁰ BALDWIN & CLARK, *supra* note 11.

⁵¹ *Id.*

through these costs of complexity.⁵² (The name 360 referred, of course, to all points of the compass—one system reconfigurable for all needs.) But the 360 was very much a proprietary system, enclosed within the silo walls of IBM, which was integrated vertically into all stages of computer production, from semiconductor fabrication through to sales and service. Here again, the option value of the system attracted entry on modules, often in the form of spin-offs by IBM insiders who knew the system and interfaces well and who, as specialists, could also see ways of improving the modules. These interlopers were the so-called “plug-compatible” competitors. IBM responded by lowering prices, changing its licensing arrangements, and, significantly, attempting to make its newer 370 System less modular.⁵³

The plug-compatible makers responded with private antitrust suits, and the Department of Justice joined in with a suit that lasted more than ten years and was finally withdrawn.⁵⁴ But, as we have seen, it was not antitrust but rather competition from an entirely new modular system—the PC—that led to the Schumpeterian transformation of the computer industry. By placing its imprimatur on a variant of an existing hobbyist standard in 1981, IBM was an important part of the birth of this new modular system. Could the company have controlled the new platform the way it had controlled the 360/370 System? IBM was certainly careless in asserting control over this new standard, though it is important to keep in mind that no one in 1981 foresaw anything like the potential the PC platform would unleash. But in hindsight, it seems clear even the best attempts of any single company to control the personal computer would ultimately have failed.⁵⁵ With the invention of the integrated circuit, the trajectory of steadily increasing density of transistors on a chip—Moore’s Law⁵⁶—came to drive modular innovation; and improvement in the power of microprocessors, memory, and other integrated-circuit components rapidly dwarfed any advantages IBM once had in circuit design, peripherals, and marketing.

⁵² EMERSON W. PUGH, *BUILDING IBM: SHAPING AN INDUSTRY AND ITS TECHNOLOGY* (Cambridge Univ. Press 1991).

⁵³ Gerald W. Brock, *Dominant Firm Response to Competitive Challenge: Peripheral Equipment Manufacturers’ Suits Against IBM*, in *THE ANTITRUST REVOLUTION* 160, 163-68 (John E. Kwoka & Lawrence J. White eds., 1st ed. 1989).

⁵⁴ *Id.* at 181 n.2 (“The government case was filed at the end of the Johnson administration, was in preparation throughout the Nixon administration, began trial in the Ford administration, continued trial throughout the Carter administration, and was abandoned in the Reagan administration, thirteen years after it began.”); *see generally* *Transamerica Computer Co. v. IBM Corp.*, 698 F.2d 1377 (9th Cir. 1983); *Cal. Computer Prods. v. IBM Corp.*, 613 F.2d 727 (9th Cir. 1979); *Telex Corp. v. IBM Corp.*, 367 F. Supp. 258 (N.D. Okla. 1973), *rev’d*, 510 F.2d 894 (10th Cir. 1975).

⁵⁵ Apple survived many lean years as a niche player, and the company’s present-day success is not the result of its PC business. Moreover, however distinctive the Macintosh may be from the Wintel standard, it is arguably deeply imbedded, perhaps increasingly imbedded, in the same ecosystem, to the point that today a Mac is an Intel-based system perfectly capable of running Microsoft Windows.

⁵⁶ Gordon Moore, *Cramming More Components onto Integrated Circuits*, *ELECTRONICS*, Apr. 1965 at 114.

By the end of the century, the computer industry had become a modular cluster. As Baldwin and Clark and others have pointed out, the companies best able to earn rents in such a cluster were not those who attempted to contain the platform but rather those who were best able to take advantage of the unleashed option value of the modular system.⁵⁷ These included players like Microsoft and Intel, who controlled key bottlenecks—Microsoft through its control (via copyright and secrecy) of the operating system standards and Intel through its dominance (via first-mover advantages) of the microprocessor standards. But it also included Dell, which succeeded not by controlling a bottleneck but by recognizing more quickly the value of modularity and leveraging it more thoroughly than others—in effect, *stopping* attempts to bottle up the option value of the system quicker than others.

In the first decade of the twenty-first century, the PC modular cluster is feeling heat from new platforms based around handheld devices. Here, Google stands in the center of the Android ecosystem. As we saw, however, Google chooses not to earn rents by controlling the bottleneck of the operating system, permitting affiliated consumers, hardware makers, and app developers to interact relatively freely.⁵⁸ Instead, Google earns rents because affiliation with Android encourages affiliation with its panoply of internet-based services and the advertising and other revenue those services pull in. By contrast, Apple retains control over the hardware for its iOS platform while encouraging (supervised) interaction between affiliated consumers and developers. Compared to Android, Apple's iOS is thus a "walled garden."⁵⁹ But the centralized control this approach affords is essential to a company that depends on its rents for continued superiority in platform design and integration, rather than on affiliation with a larger rent-generating ecosystem. As of this writing, both the Android and Apple platforms appear to be thriving, and have driven out of existence—or will soon drive out—alternative platforms like Palm WebOS, Nokia Symbian, and perhaps Blackberry. Microsoft remains a smaller player with its Windows Phone, though it has the potential to tap into its sources of rents in the PC world by integrating future mobile and PC operating systems.

One striking result of the digital platform competition of the last few decades is that the cutting edge, if not perhaps the center of gravity, of the electronics industry has returned to the United States. Apple is not only the

⁵⁷ See generally BALDWIN & CLARK, *supra* note 11.

⁵⁸ This is true despite the fact that Google purchased a hardware maker, Motorola Mobility. It is likely that Google was more interested in the value of Motorola's portfolio of patents in the battle for overall ecosystem rents (*cf.* our discussion of RCA above) than it was in using its ownership of Motorola to control and earn rents from Android hardware. Similarly, Microsoft has forged an alliance with Nokia, though it is not yet clear whether Microsoft expects the alliance to yield design rents in the manner of Apple.

⁵⁹ Salil K. Mehra, *Paradise is a Walled Garden?*, 18 GEO. MASON L. REV. 889, 889-92 (2011).

largest electronics firm in the world, it is the largest *consumer* electronics firm in the world. This turn of events would have startled observers in the 1980s, who, having witnessed the demise of American consumer electronics, were worried that semiconductors and the rest of electronics generally would soon also disappear.⁶⁰ Most of these same observers were sure that the success of Japan lay in the high level of integration of its electronics firms, not to mention the advantages those firms received from close contact with wise industrial-planning agencies. By contrast, most felt the American industry was far too “fragmented.” In the event, of course, fragmentation proved to be exactly the right structure to tap the option value of modular platforms.

Does this analysis imply that modular platforms will always beat approaches that are closed, non-modular, or proprietary? In the short and medium runs, this is certainly not true. Institutions, including intellectual property rights, can certainly favor closed systems, as can strategic vectors. From a wider perspective, however, one can argue that the option value of an open system always wins. In a sense, this is the force behind Adam Smith’s famous saying that the division of labor is limited by the extent of the market: as the extent of the market grows, a finer division of labor—especially a division of labor across markets rather than within firms⁶¹—is able to generate faster trial-and-error learning and to take advantage of something like the option value of modularity. In the end, this is an idea not far removed from Ferguson’s conceit of “killer apps.” Societies prosper when their institutions encourage individuals to take advantage of the widely dispersed bits of knowledge each possesses while at the same time coordinating those bits within coherent platforms.

⁶⁰ Richard M. Langlois & W. Edward Steinmueller, *Strategy and Circumstance: The Response of American Firms to Japanese Competition in Semiconductors*, 21 *STRAT. MGMT. J.* 1163, 1163 (2000) (“Shrill cries arose from the literature of public policy, warning that the American semiconductor industry would soon share the fate of the lamented American consumer electronics business.”).

⁶¹ See Richard R. Nelson & Sidney G. Winter, *In Search of Useful Theory of Innovation*, 6 *RES. POL’Y* 36, 73 (1977) (arguing that, at the technological frontier, innovation occurs most rapidly under an institutional structure in which many players are engaged in technological and market competition).



PLATFORMS, TEAMWORK AND CREATIVITY:
MEDIATING HIERARCHS IN THE NEW ECONOMY

*Salil Mehra**

ABSTRACT

This paper argues that the concept of team production (Blair and Stout), while it has failed to supplant shareholder primacy as a positive theory of understanding corporate organization and behavior, in fact explains steps taken by several successful platform operators that seek to encourage complementary creativity by *external* team members. Examples of these steps include Wikipedia's delegation of important roles in preserving community quality to governance and dispute resolution bodies, and Mark Zuckerberg's cautionary letter to investors in connection with Facebook's IPO. The team production model hinges on the delegation of authority to a relatively reputable mediating hierarchy, or similar body, that team members trust *ex ante* to do rough justice *ex post* in allocating rewards from collective activity, particularly where individual contributions and their returns are difficult to disaggregate and affix values to. Where litigation or money-back guarantees are unlikely to work, these platforms seek to use the *ex ante* reputation of the hierarchy, or the hierarchy's commitment to delegate to representative institutions, to commit to an *ex post* allocation of returns that induces user reliance. The use of mediating hierarchy strategies suggests that, rather than applying "net neutrality" or essential facilities mandates for openness, in these circumstances law might provide institutions or rules that foster this type of strategy.

I. INTRODUCTION

In February 2012, Facebook filed papers for its long-awaited initial public offering; the press was primarily captivated by the massive \$5 billion flotation, and the lofty valuation it suggested for the corporation as a whole. A second focus, though, was on a letter from Facebook's founder Mark Zuckerberg that accompanied the offering. The letter cautioned potential investors that they were dealing with a different animal than they were used to:

* James E. Beasley Professor of Law, Temple University, Beasley School of Law. sme-hra@temple.edu.

As I said above, Facebook was not originally founded to be a company. We've always cared primarily about our social mission, the services we're building and the people who use them. This is a different approach for a public company to take, so I want to explain why I think it works.

....

Simply put: we don't build services to make money; we make money to build better services.

And we think this is a good way to build something. These days I think more and more people want to use services from companies that believe in something beyond simply maximizing profits.¹

There are several reasons to set forward such cautions, not least of which is the desire to seek patience in case returns are not what they seem. However, given his retention of a majority equity stake—bigger than what Bill Gates retained at the time of Microsoft's IPO—he should have little fear of ouster *a la* Steve Jobs at “Apple 1.0.”² And despite the strength of shareholder primacy as a guiding principle of corporate governance, today's Delaware courts are unlikely to override a board of directors' decision to, for example, reinvest earnings into the company for long-term growth as opposed to paying out dividends in the short term.³

So why write such a letter? One possibility is that it is a statement to “the team.” Some media reports focused on the letter's endorsement of Facebook's continued use of “hacker way.” In particular, this was explained with reference to the internal all-night “hackathons” and general willingness to consider newly generated internal improvements even by fresh hires.

But the “team” need not be so narrowly defined. As others have noted, Facebook's value resides not simply in unique software code, but in the complementary content—think of photos and news from your friends and relatives—and innovation—think Zynga and FarmVille—that users and independent developers create. This creativity makes the network more valuable, but, ultimately, these *external* members of Facebook's “team”

¹ Facebook Inc., Registration Statement (Form S-1) 67-70 (Feb. 1, 2012), available at <http://sec.gov/Archives/edgar/data/1326801/000119312512034517/d287954ds1.htm>.

² Steven M. Davidoff, *A Big Bet on Zuckerberg*, N.Y. TIMES (Feb. 2, 2012), <http://dealbook.nytimes.com/2012/02/02/a-big-bet-on-zuckerberg> (calculating that Zuckerberg will control at least 57.1% of the voting shares initially).

³ See *Dodge v. Ford*, 170 N.W. 668, 684-85 (Mich. 1919) (ordering board dominated by Henry Ford to pay out earnings as dividends to the Dodge Brothers—shareholders who were starting their own car company—where he was accused of running Ford Motor Company as a “semi-eleemosynary” enterprise). The case is rarely cited, and arguably bad law, but familiar to virtually every American law student who takes Corporations. See generally Lynn A. Stout, *Why We Should Stop Teaching Dodge v. Ford*, 3 VA. L. & BUS. REV. 163 (2008) (pointing out these defects with the case).

rely on the platform operator's continued cooperation. These opportunistic policy changes by Facebook could destroy their investments' value in complementary content and innovation.⁴

In this brief Article, produced in connection with a symposium hosted by the *Journal of Law, Economics & Policy*, I argue that there is another way to understand Zuckerberg's letter—or Wikipedia's attempts to build an ethos and governance institution,⁵ or Apple's ability to foster a community identity among hundreds of millions of customers.⁶ The concept of team production, as introduced by Professors Margaret Blair and Lynn Stout, has largely failed to supplant shareholder primacy as a positive theory of understanding corporate organization and behavior. However, Blair and Stout's suggesting the need for a "mediating hierarch" to divvy up returns when it is difficult to disaggregate team members' contributions appears to fit the behavior of platforms that seek to encourage complementary creativity by *external* team members. Where *ex ante* contracts⁷ or *ex post* legal remedies⁸ are unlikely to work, these platforms seek to use the *ex ante* reputation of the hierarch, or the hierarch's commitment to delegate to representative institutions, to commit to an *ex post* allocation of returns that induces user reliance.

Section I explains why a simple static two-sided market approach may not capture important key platform dynamics, particularly the possibility of opportunism by the platform operator and quality sensitivity by the users. Section II discusses the considerations that make other pre-commitment

⁴ See Salil Mehra, *Paradise is a Walled Garden? Trust, Antitrust and User Dynamism*, 18 GEO. MASON L. REV. 889 (2011) (arguing that opportunism by a platform host who induces reliance in a dynamic environment may justify some forms of regulation or intervention); see also Jonathan Barnett, *The Host's Dilemma: Strategic Forfeiture in Platform Markets for Informational Goods*, 124 HARV. L. REV. 1861 (2011) (arguing that platform hosts can strategically forfeit rights to intellectual property as a way of guaranteeing cooperation with those who use the platform).

⁵ See David Hoffman & Salil Mehra, *Wikitruth Through Wikiorder*, 59 EMORY L.J. 151 (2009) (describing and analyzing creation of dispute resolution institutions and their interplay with norms concerning authorship and participation in Wikipedia); JOSEPH MICHAEL REAGLE, *GOOD FAITH COLLABORATION: THE CULTURE OF WIKIPEDIA* (2010) (discussing conscious fostering of a "good faith collaboration culture").

⁶ See Steve Jobs: *The Magician*, THE ECONOMIST (Oct. 8, 2011), <http://www.economist.com/node/21531529> (observing that "Apple users feel themselves to be part of a community, with Mr[.] Jobs as its leader").

⁷ See Jenna Wortham & Nick Wingfield, *Microsoft is Writing Checks to Fill out Its App Store*, THE NEW YORK TIMES (Apr. 5, 2012), <http://www.nytimes.com/2012/04/06/technology/to-fill-out-its-app-store-microsoft-wields-its-checkbook.html?pagewanted=all> (describing Microsoft's paying or financing developers to make Windows Phone versions of smartphone applications – with limited success in comparison with Apple or Google, which "do not have to pay developers").

⁸ See *Tasini v. AOL Inc. et al.*, 851 F. Supp. 2d 734 (S.D.N.Y. 2012) (dismissing claim of 9000-member "Huffington Union of Bloggers" for one-third of \$105 million sales price of the Huffington Post based on value of their contributions to the development of the platform for politically left-leaning news and opinion).

forms difficult and that lead to the mediating hierarch approach. Section III suggests some legal and policy implications and is then followed by a brief conclusion.

II. WHY A TWO-SIDED MARKET MODEL MAY FIT VISA AND WINDOWS BUT NOT FACEBOOK

In a remarkably short time period, the two-sided market model—or, as David Evans has rightly suggested would be a better name, the “two-sided platform” model⁹—became popular in the literature surrounding industries characterized by network effects.¹⁰ That attention is well-deserved; the model provides a powerful, positive explanation of real-world strategies surrounding products such as operating systems¹¹ and credit cards.¹² The basic intuition is well-known to those familiar with the online world: where

⁹ See David S. Evans, *Two-Sided Markets*, in PLATFORM ECONOMICS: ESSAYS ON MULTI-SIDED BUSINESSES 135, 137 (2011) [hereinafter PLATFORM ECONOMICS].

¹⁰ This powerful model may not fit every platform with two sides. In particular, there may be a failure of fit where the same participants act on both sides as both producers/creators and consumers/users. The two-sided market approach could, of course, still work well as a model if participants' activity could be segmented neatly—they might experience one “price” as a producer/creator on one side of a platform, and another as a consumer/user on the other side. Think, for example, of a sandwich shop proprietor who accepts Visa cards as payment, but who also uses Visa to purchase supplies at her local warehouse club. She would pay a different fee on the merchant side than on the consumer side; Visa would set the two fees in order to maximize overall profit. The two-sided market concept may fit more poorly if participation cannot be so neatly segmented. That may be because the activity as creator and consumer are intertwined so that exit from one side of the market means exit from both sides. It might be argued that this just describes one-sided markets with network effects, like having your telephone disconnected. But exiting Facebook on the content sharing side while hoping to continue enjoying the content consumption side might be like hoping to have a telephone that you can hear other people on while they cannot hear you—a situation that is unlikely to be very stable. Of course, Facebook has more than two sides, as it possesses not only the content sharing and content consumption sides, but also a market for apps, especially games by the Zynga corporation, and its latent internal capacity to mine the wealth of user data it possesses in order to develop “social search” or other future capacities. See, e.g., Eric Jackson, *Facebook's Yahoo Patent Problem*, CNBC STOCK BLOG (Jan. 31, 2012, 12:36 PM), http://www.cnbc.com/id/46204987/Facebook_s_Yahoo_Patent_Problem (describing how Facebook and its competitors are assembling patents, including by purchasing from defunct social network predecessors, to gear up for future litigation over search capacity based on social connections).

¹¹ Geoffrey Parker & William Van Alstyne, *Information Complements, Substitutes, and Strategic Product Design* (William Davidson Working Paper Series, Paper No. 299), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=249585 (explaining how insights of Katz and Shapiro concerning network effects in a one-sided market such as telephony intersect with price discrimination and product differentiation to create incentives to, for example, give away products in one market to shift the demand curve and increase sales in another market).

¹² Jean Charles Rochet & Jean Tirole, *Platform Competition in Two-Sided Markets*, 1 J. EUR. ECON. ASS'N 990 (2003), available at <http://www.rchss.sinica.edu.tw/cibs/pdf/RochetTirole3.pdf> (observing, similarly to Parker & Van Alstyne, *supra* note 11, incentives for platform operator's reduced pricing in one market to stimulate demand in another market).

a platform seeks to attract viewers' eyeballs and advertisers' dollars, for example, it should not rationally try to maximize profit independently in each market, e.g., eyeballs vs. advertisers. Instead, it may make sense to charge a reduced or zero price in one market in order to increase demand in the other.¹³ This intuition explains the prevalence of distribution at zero monetary cost on platforms characterized by two-sided markets.¹⁴

The wrinkle is that over time, the platform operator may shift the relative prices and costs between the two sides of the market. This possibility is essentially a variant on the concern that platform operators may start as "open" platforms, only to close themselves into walled gardens once they have built up a sufficient installed base.¹⁵ The dichotomy between a wholly open platform and a completely "closed" walled garden is almost certainly a false one. In reality, even open platforms have their gatekeepers.¹⁶ Because open platforms and closed walled gardens are not the only possible outcomes, but are instead endpoints on a spectrum, the possibility exists that platform operators may subsequently shift the relative prices between sides of the market. For example, a platform that offers the user side a relatively good deal now in order to push out the demand curve on the application (app) developer/advertiser side may find that, in the future, they would like to change the relative prices or quality—terms—that users experience. Whether that should be a consideration for courts or regulators depends on factors such as deception or imperfect information, which will be discussed in the succeeding sections.

Again, consider Facebook, which currently has a user side and an app developer/advertiser side. Users enjoy zero monetary cost access to Facebook, which increases demand on the app developer/advertiser side. But as many commentators have observed, Facebook is also compiling large

¹³ *Id.*

¹⁴ See, e.g., CHRIS ANDERSON, *FREE: THE FUTURE OF A RADICAL PRICE* (2009) (noting many examples of such marketing/distribution).

¹⁵ See TIM WU, *HEARING ON DIGITAL ECONOMY: OVERSIGHT OF INNOVATION CATALYSTS 6* (2012), available at <http://www.oecd.org/regreform/liberalisationandcompetitioninterventioninregulatedsectors/49623337.pdf> (describing problem of platforms that declare themselves to be open at the outset, then later change policy and the resulting "corrupting effects on the entire system of platform-based innovation"); see also Barnett, *supra* note 4 (arguing that the distinction between open and closed platforms is overemphasized, since both operate under the same insolvency constraints, so "open" platforms must find some closed areas in which to recoup their investments, and "closed" platforms have methods to guarantee degrees of openness to others); Mehra, *supra* note 4 (arguing that disclosure and enforcement of reliance interests may address these concerns so that walled gardens, with governance, may prove superior to open platforms); Joshua Wright & Geoffrey Manne, *The Case Against the Case Against Google*, 34 HARV. J.L. & PUB. POL'Y 171 (2011).

¹⁶ See Hoffman & Mehra, *supra* note 5, at 195-96 (discussing how Wikipedia, the poster child for open mass collaboration, nevertheless has institutions aimed at weeding out problematic users); see also Mehra, *supra* note 4, at 901 (discussing how Wikipedia controversially "transwikis" material deemed not fit for Wikipedia—but that draws traffic—such as niche pop culture-related entries, to an affiliated, for-profit, less open site).

amounts of data that it may seek to monetize in the future through data mining or social search.¹⁷ This represents a potential third side of the platform, even if it currently only exists, as Justice Scalia says in *Trinko* “deep within the bowels of” Facebook.¹⁸ Incidentally, the fact that many analysts think that financially exploiting Facebook’s data mining and social search capabilities is the platform’s most valuable future strategy may highlight the difficulty that antitrust has with the possibility that development of products or even markets may be a function of regulation—or lack thereof—itsself. Scalia, in *Trinko*, seems to suggest that markets that exist only in gestation within a host firm may avoid antitrust scrutiny from courts.¹⁹ Similarly, any antitrust consideration of Facebook’s ability to transform its internal data capacity into a business or market for social search or transformative behavioral marketing will also need to confront the impact of other regulators²⁰—even if that means one segment of the FTC considering the activity of another.

Development of data mining/social search could result in an *ex post* change to the prices that the user side and the app/advertiser side confront *ex ante*. In the absence of disclosure and pre-commitment, users might rightly fear that platform operators such as Facebook might face a strong incentive to attract users with zero-cost access in the adoption stage, only to change the prices, costs, or quality that users experience after large-scale adoption yields a dominant position. Where lock-in, through lack of data portability or otherwise, hinders exit, such a strategy could rationally increase profit despite deterring new users.²¹

Another complication is that users who choose between platforms with zero pecuniary cost are likely to focus on quality when making choices. As the economist Albert Hirschmann described four decades ago, clubs, congregations, schools, and even whole societies share this outlook in which

¹⁷ See David S. Evans, *The Online Advertising Industry: Economics, Evolution, and Privacy*, in PLATFORM ECONOMICS: ESSAYS ON MULTI-SIDED BUSINESSES 226, 250 (2011) (discussing this problem in the context of a two-sided platform). The possibility that platform operators may make this shift in a way that exploits consumers’ imperfect information raises a number of concerns including privacy issues.

¹⁸ Verizon Commc’n Inc. v. Law Offices of Curtis V. Trinko, 540 U.S. 398, 410 (2004).

¹⁹ Of course, even if they did count, it might be very difficult for courts to figure out how to treat them. As the now-withdrawn Section 2 report noted, siding with the Court’s opinion in *Trinko*, “judges and enforcement agencies are ill-equipped to set and supervise the terms on which inputs, property rights, or resources are provided.” U.S. DEP’T OF JUSTICE, COMPETITION AND MONOPOLY: SINGLE FIRM CONDUCT UNDER SECTION TWO OF THE SHERMAN ACT (2008), available at <http://www.justice.gov/atr/public/reports/236681.pdf>.

²⁰ See Evans, *supra* note 17 (discussing privacy concerns in the context of social networks as two-sided platforms).

²¹ See Eastman Kodak Co. v. Image Tech. Servs., Inc., 504 U.S. 451, 476-77 (1992).

changes in quality are particularly salient.²² As Hirschmann recognized, participants in such organizations are both consumers and producers at the same time, and their exit from one side usually accompanies an exit from the other.²³ Furthermore, as Hirschmann observed, if participants are quality-sensitive, if quality sensitivity among the participants is heterogeneous, and if quality deteriorates when quality sensitive participants exit, then the start of a quality decline can lead to a downward spiral. Those who exit first cause quality degradation that may be sufficient to lead others to “jump ship,” and so on. In particular, Hirschmann contrasted this susceptibility with the normal tendency of markets to self-correct; a seller whose product quality deteriorates faces a strong incentive to cut price so that the quality-price bundle remains competitive in the market.²⁴ I have previously argued that this possibility provides an alternative, though not mutually exclusive, explanation to low entry barriers for why platform dominance can seem so fragile—one such example is how quickly Myspace was overtaken.²⁵

Social networking fits particularly well into Hirschmann’s paradigm. While Facebook garners much of the attention, similar but smaller platforms also fit the commonly-used definition for social networking as a service that connects users within a bounded system, enabling them to make new connections derived from their existing ones.²⁶ These communities include not only general social networking sites such as Google+, LinkedIn, and the aforementioned Myspace, but also specialized sites that focus on identity, personal interests, or causes.²⁷ The community-based nature of such platforms may make them susceptible to the quality-decline spirals that Hirschmann posited.

The subset of two-sided markets where *ex post* opportunism is possible and concerns about quality sensitivity dominate may delineate some platforms that foster broader communication, creativity, and innovation. As the next section suggests, the mediating hierarch concept may describe a strategy for platform operators to employ in order to increase user trust so that users will not be victims of opportunism and so quality will not decline.

²² See ALBERT HIRSCHMANN, EXIT, VOICE, AND LOYALTY 48 (1970) (describing importance of variance in response to quality decline in some organizations, and how this variance leads to a different result than an increase in price otherwise would).

²³ *Id.* at 102 (describing markets in which individuals participate both as consumer and producer).

²⁴ *Id.*

²⁵ See Mehra, *supra* note 4, at 910-11.

²⁶ Danah M. Boyd & Nicole Ellision, *Social Network Sites: Definition, History, and Scholarship*, 13 J. COMPUTER-MEDIATED COMM. 210 (2007) (defining social networking as “[w]eb-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connections, and (3) view and traverse their list of connections and those made by others within the system”); see also Spencer Waller, *Antitrust and Social Networking*, 90 N.C. L. REV. 1771, 1777 (2012).

²⁷ See Waller, *supra* note 26, at 1776-78 (describing the range of social networking).

III. THE MEDIATING HIERARCH ONLINE

In the AT&T and Microsoft-case studies that dominate the earlier literature on network effects and antitrust, the assumption was often that the investment was most often made by the network operator, and sometimes, as in the Microsoft case, by the software side of a two-sided platform. But in more recently-developed platforms, users often make significant investments in the platform as well. Now, both users and platform operators share a common dilemma: they both make commitments to the platform that they may not later be able to fully recover.²⁸

The problem of firm-specific investment is not a new one. Indeed, a central issue in corporate law is finding the best solution for the fairly common situation where parties make firm-specific investments that are difficult to monitor, measure, and reduce to explicit contracts.²⁹ This problem is difficult enough in the general corporate context; it becomes fiendishly more difficult when considering a technologically dynamic platform. As Thomas Hazlett, David Teece, and Leonard Waverman have noted, even among closely vertically-related firms in the smartphone industry, disruptive innovation in handsets, software, and content have caused profits to flow away from networks.³⁰ Instead, the authors suggest that profits flow towards firms that provide complements that override the networks, in often unpredictable ways.³¹

One line of argument based on property rights suggests that, where multiple parties' investments are difficult to measure and monitor, the solution may be to grant control of the enterprise to the party whose firm-specific investment is the most crucial.³² However, as Raghuram Rajan and Luigi Zingales noted, granting ownership or control to one party does not ensure optimal investment.³³ This is due to the decision the party must make between either making the firm-specific investment or capturing a share of rents from the firm by simply selling control to a third party—possibly by profiting from the non-controlling party's firm-specific invest-

²⁸ *Id.* at 1788 (describing difficulty of data portability from Facebook to elsewhere).

²⁹ See Margaret M. Blair & Lynn A. Stout, *A Team Production Theory of Corporate Law*, 85 VA. L. REV. 247, 272-73 (1999) (discussing leading arguments concerning this problem).

³⁰ See Thomas Hazlett, David Teece & Leonard Waverman, *Walled Garden Rivalry: The Creation of Mobile Network Ecosystems* 5-16 (George Mason Law & Economics Research, Working Paper No. 11-50, 2011), available at <http://ssrn.com/abstract=1963427>.

³¹ *Id.*

³² Oliver Hart & John Moore, *Property Rights and the Nature of the Firm*, 98 J. POL. ECON. 1119, 1149 (1990); Jonathan Macey, *Externalities, Firm-Specific Capital Investments, and the Legal Treatment of Fundamental Corporate Changes*, 1989 DUKE L.J. 173, 174-78 (1989) (arguing for shareholder primacy on the grounds that other stakeholders in a corporation can protect themselves more easily through contract).

³³ See Raghuram Rajan & Luigi Zingales, *Power in the Theory of a Firm*, 113 Q. J. ECON. 387, 395 (1998).

ments.³⁴ The two-player game below illustrates this problem in an oversimplified way. In the game, if both the platform operator and the user jointly and cooperatively invest, then they wind up in the southeast quadrant where the joint payoff is the highest. But the operator does worse (-1) if she cooperatively invests in a platform and the user does not (0) than if no investment is made, in which case each party receives 0. The operator does best if the user invests, but the operator, instead of cooperatively investing, appropriates the user's investment and sells to a third party. The operator has a dominant strategy of not cooperatively investing. Due to this, the user should decide not to invest, thus creating a jointly inferior result by landing both in the northwest quadrant.

		Operator (chooses column)	
		Do not cooperate	Cooperate
User (chooses row)	Do not invest	0	-1
	Invest	0	0
		2	1
		-1	1

Of course, this is an oversimplification. The relationship between the user and the platform operator is not a one-shot game, and the actual number of players may be quite large. In an iterated game, particularly one without a fixed end, trust and cooperation may build up between the players. However, the possibility remains that the user may fear the operator's decision to exploit the user in an end game strategy. Furthermore, barring a method of pre-commitment, that concern may cause the game to "zipper" back to a jointly inferior result.

Again, taking Facebook or Wikipedia as examples, the massive user bases of each platform have contributed a tremendous amount of content, while a smaller set of developers have generated applications—Facebook—or code improvements—Wikipedia. In each platform, those who have made such firm-specific investments might fear that the platform operator might find that, rather than making its own jointly optimal firm-specific investments, selling control to a third party might be individually optimal, even if jointly inferior. For Facebook's users, that concern might be the

³⁴ *Id.*

buyers in an IPO; the specter that has haunted Wikipedia's users in the past has been a potential move towards advertising and a for-profit model.

This fear is not purely academic; open platforms have seen *ex post* opportunistic appropriation through policy changes in the past,³⁵ potentially creating future user mistrust of similar enterprises.³⁶ *Ex ante* contracts have their limitations; Microsoft has attempted to pay or provide financing to developers in advance to create applications for its Windows Phone operating system, but given its small market share—and perhaps the legacy of Microsoft's past appropriation of applications developers' innovation³⁷—these efforts have yet to succeed. Similarly, *ex post* legal remedies may not provide much redress for contributors to the value of a platform. The 9000-user Huffington Union of Bloggers learned this when their asserted “unjust enrichment”-based claim to a share of the sale price of the Huffington Post was denied on traditional freedom of contract grounds.³⁸

As Margaret Blair and Lynn Stout have argued, the solution to the problem of opportunism and joint firm-specific investment lies in the possibility of a mediating hierarch:

[I]ndividuals will only want to be part of a team if by doing so they can share in the economic surplus generated by team production. . . . [T]eam members intuitively understand that it will be difficult to convince others to invest firm-specific resources in team production if shirking and rent-seeking go uncontrolled. Thus, they realize that it is in their own self-interest to create a higher authority—a hierarch—that can limit shirking and deter rent-seeking behavior among team members. In other words, team members submit to hierarchy not for the hierarch's benefit, but for their own.³⁹

Blair and Stout presented this positive theory as an alternative to the shareholder primacy model of the corporation. While it has had limited success in that endeavor, the dynamic it discusses and the mechanism it describes fit observable platform behavior quite well.

³⁵ See Mehra, *supra* note 4, at 896 (discussing user concern surrounding several examples, including CDDB/Gracenote and Wikipedia).

³⁶ See Wu, *supra* note 15, at 5 (expressing general concern about the corrupting effect of opportunistic policy changes on platform-based innovation).

³⁷ See Wortham & Wingfield, *supra* note 7; see also *United States v. Microsoft Corp.*, 253 F.3d 34 (D.C. Cir. 2001) (en banc) (describing Microsoft's attempts to use tying and bundling to supplant the innovative browser application Netscape); *Sun Microsystems v. Microsoft Corp.*, EC Comm 1 (Comp/C-3/37.792) (March 24, 2004), available at <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32007D0053:EN:NOT#top> (European Commission decision finding that Microsoft abused its dominant position by tying its Windows Media Player with the Windows OS by technically integrating WMP software into Windows, steps that supplanted the innovative audiovisual application RealPlayer).

³⁸ See *Tasini v. AOL Inc. et al.*, 851 F. Supp. 2d 734, 740 (S.D.N.Y. 2012) (order dismissing complaint based on implied contract between platform and users under which court concluded that “the plaintiffs submitted their materials to The Huffington Post with no expectation of monetary compensation and that they got what they paid for—exposure in The Huffington Post”).

³⁹ Blair & Stout, *supra* note 29, at 274.

Viewed through the lens of team production, the Zuckerberg letter may make sense as an attempt to fill the role of mediating hierarch in an effort to reassure those who have made or continue to make firm-specific investments. Whether such a statement can succeed in doing so remains to be seen. Wikipedia has addressed concerns about the possibility of advertising and the “transwikiing” user-generated content to an affiliated for-profit site by adopting lenient content-portability policies and building institutions that spread governance authority among a portion of its user base.⁴⁰ Both approaches may have merit, though Wikipedia’s portability policy is in part a commitment not to proceed in the direction of a walled garden. As the next section discusses, these attempts to deal with the problems of team production have related policy implications.

IV. IMPLICATIONS

Almost all scholars who have considered the question of how regulation should address dominant platforms have expressed caution, particularly where antitrust law is concerned; nevertheless, the degree of caution has varied significantly.⁴¹ Because antitrust law dealing with dominant platforms invariably involves the law of monopolization—a contested and somewhat unstable area—uncertainty is unavoidable. Even worse, the controversial subset of monopolization law concerning essential facilities, mandated interconnection, and compulsory licensing is also implicated. Of course, these areas inevitably overlap with discussions of network neutrality.⁴²

Modeling dominant platforms through team production raises a somewhat different set of issues. While these issues may not obviate the anti-trust discussion, policy steps that foster solutions to dominant platforms’ team production dilemma may, in the long run, reduce the need for anti-

⁴⁰ See Jon Bernstein, *Wikipedia’s Benevolent Dictator*, NEW STATESMAN (Feb. 3, 2011), <http://www.newstatesman.com/digital/2011/01/jimmy-wales-wikipedia-site> (explaining that the absence of commercialism on the site fosters trust from dynamic users who are interested only in “sharing the passion”); Ivan Beschastnikh, et al., *Wikipedian Self-Governance in Action: Motivating the Policy Lens*, in PROCEEDINGS OF THE SECOND INTERNATIONAL CONFERENCE ON BLOGS AND SOCIAL MEDIA 27, 29 (2008), available at <http://www.aai.org/Papers/ICWSM/2008/ICWSM08-011.pdf>; Andrea Forte & Amy Bruckman, *Scaling Consensus: Increasing Decentralization in Wikipedia Governance*, PROCEEDINGS OF THE 41ST ANNUAL HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES 4 (2008), <http://www.computer.org/portal/web/csdl/doi/10.1109/HICSS.2008.383>; Andrea Forte & Amy Bruckman, *Why Do People Write for Wikipedia? Incentives to Contribute to Open-Content Publishing*, JELLIS.ORG, <http://jellis.org/work/group2005/papers/forteBruckmanIncentivesGroup.pdf> (last visited Oct. 19, 2012); Hoffman & Mehra, *supra* note 5, at 203-04.

⁴¹ Compare Waller, *supra* note 26, with Wright & Manne, *supra* note 15, at 213. See also Evans, *supra* note 17, at 262. Of course, these commentators are not always focused on the same platforms.

⁴² See, e.g., Spencer Waller & Brett Frischmann, *Revitalizing Essential Facilities*, 75 ANTITRUST L.J. 1, 26 (2008).

trust-based regulatory intervention. Thinking through the implications of the mediating hierarch concept for dominant platforms could lead to conclusions that do not necessarily make easy ideological bedfellows.

First, as suggested by the Zuckerberg letter example, the team-production/mediating hierarch concept implies that norms of corporate governance should allow directors or officers of dominant platforms the latitude to balance the demands of users versus those of shareholders. Such a conclusion naturally follows from the team-production model's portrayal of both, making investments that are difficult to disaggregate in explaining the platform's success. As a result, the law may need to recognize that shareholder advocates' arguments are relatively weaker in the context of some platforms.

Second, to the extent that platform operators are engaged in a difficult-to-formalize contractual relationship with users, concerns about opportunism may suggest a greater role for consumer protection-based intervention. Particularly where aggregate litigation is also likely to be a weak alternative, the possibility of a platform life cycle ending in opportunistic appropriation by the platform operator may tend to corrupt the user-platform joint production model in ways that could be harmful to innovation.⁴³ The nature of such regulation—whether compelled disclosure, enforcement of commitments, or encouragement of industry norm generation—could be very tricky. The mediating hierarch's decisions will be subject to changes in relative economic and political power among the stakeholders,⁴⁴ making the regulatory challenge that much more complicated.

Finally, application of the mediating hierarch model to user-platform operator joint production may suggest the need for supporting legal rules and forms. As Blair and Stout recognized in initially setting forth their model, the hierarch's trustworthiness is critical to their ability to foster team production.⁴⁵ The Blair and Stout project was essentially a positive theory explaining the corporation as a longstanding entity. User-platform joint production is at its incipiency. As a result, generating rules and forms that bolster trust in the hierarch may be critical for such platforms to reach their full potential.

CONCLUSION

The nature of joint production between users and platform operators is still somewhat uncertain. Despite that uncertainty, these platforms have

⁴³ See WU, *supra* note 15.

⁴⁴ See Blair & Stout, *supra* note 29, at 325-27.

⁴⁵ Margaret M. Blair & Lynn A. Stout, *Trust, Trustworthiness, and the Behavioral Foundations of Corporate Law* 5 (Bus., Econ., and Regulatory Law, Working Paper No. 241403, 2000), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=241403.

already shown that they may be important forces of innovation and production in the future. Just as the law adapted itself to corporate form in the industrial age, and to more recent unincorporated entities in the post-industrial era, user-platform joint production may demand shifts in legal rules and forms as well. The application of the mediating hierarch model provides a useful first look at the kinds of shifts that may be required.



THE BROTHERS GRIMM BOOK OF BUSINESS MODELS: A
SURVEY OF LITERATURE AND DEVELOPMENTS IN PATENT
ACQUISITION AND LITIGATION

*Anne Layne-Farrar**

INTRODUCTION

Although they have received the lion's share of the limelight in recent years, it is not just "trolls" shaping the patent landscape today; "giants" and "dwarves" are in the thicket as well. The trolls, of course, are those patent holders that do not practice their patents, but instead seek to profit from litigating them. As trolls are relatively new to the scene, emerging in force in the 1990s, they have commanded the spotlight for some years. But the giants have always been present—those large, multinational firms with both extensive patent portfolios and products practicing them. One step down from the giants are the "dwarves"—firms with a large presence in one market entering another for which they have few or no relevant patents. The events of 2011 provided a sharp reminder to keep a watch on the giants and dwarves too—indeed on all business models involving intellectual property (IP). The patent ecosystem is a diverse one, and creatures of all stripes each add their own twist.

This article provides a survey of key patent-related events that took place in 2011 and takes stock of the current patent acquisition and litigation landscape. Specifically, I review the empirical literature on patent assertion entities (a less derogative and more descriptive term than *patent troll*, as explained below), assess the patent auctions over which the giants and dwarves battled fiercely throughout 2011, and survey the numerous patent infringement lawsuits currently underway—cases which include troll, giant, and dwarf plaintiffs. The patent acquisition and litigation sections focus on the wireless telecom ecosystem, as this is where the most heated battles are occurring. I then offer my thoughts on the forces that have led to the remarkable patent acquisition and litigation activity within wireless over the last few years. I close the article with some thoughts on the likely future of the IP battles.

* Anne Layne-Farrar is a Vice President with Charles River Associates. The author thanks Daniel Garcia Swartz, Jorge Padilla, Allan Shampine, and participants at the George Mason University School of Law conference, The Digital Inventor: How Entrepreneurs Compete on Platforms, held February 24, 2012, for helpful comments. Please send any queries or comments to alayne-farrar@compasslexecon.com.

I. PORTRAIT OF A “TROLL”

First, a bit a nomenclature. The term *patent troll* is certainly colorful and it does describe in broad terms the actions against which so many have complained: the charging of a toll for access to something the troll did not create and likely has been holding in reserve until the price was right.¹ But the term has also been used far too expansively and too pejoratively for use in constructive debate. Indeed, the term *troll* initially was applied to all non-practicing entities (NPEs); essentially any firm that was not practicing its patents was dubbed a troll.²

As has been pointed out in the academic literature, however, many NPEs do not behave like stereotypical trolls. Universities, for example, fall squarely in the NPE camp because they do not make things.³ However, university patents come from their staffs' research. Many universities have active patent licensing (or “technology transfer”) offices that seek to monetize faculty patents without “holding them in reserve,” and universities occasionally spawn commercial start-ups to further develop especially promising inventions. Universities are also not active litigators. As Colleen Chien finds in her empirical study of high-tech patent litigation, non-profit entities, of which universities are one element, account for only 1% of the patent infringement plaintiffs.⁴ Clearly universities are not what people have in mind as a patent troll, despite their non-practicing status. Indeed, universities might be better considered patent “elves,” in that they conduct meaningful research, which is then dispersed through the economy via licensing.

Research and Development (R&D) specialists offer another example of a non-practicing entity undeserving of the troll moniker.⁵ These firms concentrate on upstream research, typically spending substantial sums on R&D and innovation, but again they do not make things. To economists,

¹ The term *patent troll* was coined in the late 1990s by Peter Detkin, then-assistant general counsel at Intel but now, ironically, a managing partner at Intellectual Ventures, a firm seen by many as *the* patent troll. See discussion *infra* pp. 6-7.

² Brenda Sandburg, *You May Not Have a Choice. Trolling for Dollars*, THE RECORDER, July 30, 2001, available at <http://www.phonetel.com/pdfs/LWTrolls.pdf>.

³ See generally Mark Lemley, *Are Universities Patent Trolls?*, 18 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 609 (2008).

⁴ Colleen Chien, *Of Trolls, Davids, Goliaths, and Kings: Narratives and Evidence in the Litigation of High-Tech Patents*, 87 N.C. L. REV. 1571, 1600 (2009) (indicating in Table 3 the percentage of suits brought by type of plaintiff). A search of US Patent Office data indicates that at least 1.3% of all granted patents are assigned to entities with *university* in the title. As many universities and their associated labs do not include the word *university* in their titles, the 1.3% figure understates the proportion of patents held by universities. Moreover, the 1% figure reported by Chien includes all non-profits. Hence, it seems safe to assume that universities litigate below their share of patents.

⁵ See, e.g., Damien Geradin, Anne Layne-Farrar & Jorge Padilla, *Elves or Trolls? The Role of Nonpracticing Patent Owners in the Innovation Economy*, 21 INDUS. & CORPORATE CHANGE 1 (2011).

these entities are merely exercising their comparative advantage by focusing on what they do best—research and innovation—and ignoring what they do not do well or cannot afford to do at all—commercial production. For example, semiconductor design houses—so-called fabless chip shops that research and design integrated circuits which are then manufactured by separate fabrication plants—can fall into this category.⁶ So do many biotech firms, which emerged after the discovery of recombinant DNA technology in 1973. These biotech firms tend to focus on early-stage research and intermediate chemical inputs, leaving time-consuming and expensive drug development and commercialization to larger pharmaceutical companies.⁷ Though these are just two examples of another category of NPE elves, a careful examination would reveal R&D specialists in many different sectors throughout the economy.

Instead of profiting from the sale of goods in a downstream market, many R&D specialists profit from licensing the patents on their inventions. Because their profits often depend on a stream of licensing revenues over time, these firms tend to license broadly, diffusing technology and enabling the entry of downstream specialists—firms without meaningful patent portfolios of their own but wishing to produce goods incorporating the latest technology. Although upstream specialists can certainly be involved in patent litigation (as can any patent holder, whether practicing or not), it is important to understand the repeat-game nature of an upstream invention-based business model: significant R&D investment is followed by patenting and then by licensing, and licensing revenues can often fund the next round of R&D.

This repeated cycle can provide a counter balance to incentives for one-shot patent holdup.⁸ If firms attempt patent holdup, others in the ecosystem have strong incentives to invent around or avoid their patents in the next iteration of the product, which would eventually dry up the R&D specialists' revenues. Moreover, these firms are not charging a toll for access to something they did not create; they are charging a toll for access to something they did create and for which they are entitled to earn a return. Thus, the important distinction for this class of NPE is that licensing fees are not in and of themselves bad. Licensing fees offer a return on research investments and thus, can spur and fund more research; this is the fundamental rationale for the patent system. Rather, the complaint about trolls is

⁶ See, e.g., Bronwyn Hall & Rosemarie Ziedonis, *The Patent Paradox Revisited: An Empirical Study of Patenting in the U.S. Semiconductor Industry, 1979–1995*, 32 RAND J. ECON. 101 (2001).

⁷ Geoffrey Carr, *A Survey of the Pharmaceutical Industry: Beyond the Behemoths*, ECONOMIST, Feb. 19, 1998, available at <http://www.economist.com/node/604094>. Carr found that in 1998, roughly 18% of pharmaceutical R&D funds went toward outsourcing.

⁸ Hold-up can be a potential problem when successive innovations use the same production facilities, particularly if those innovations make earlier products obsolete. For example, an inventor of integrated circuits may work regularly with fabrication firms that have sunk significant costs into their factories.

that patent holders should be prevented from exploiting switching costs—the expenses involved when a patent holder unexpectedly appears after a licensee has made its own investments in manufacturing and commercialization that lead to patent holdup.⁹

In light of the many procompetitive, non-exploitative patent licensing arrangements that do not involve the practice of patents, the Federal Trade Commission coined the term *patent assertion entity*, or PAE, in its March 2011 report on IP, *The Evolving IP Marketplace, Aligning Patent Notice and Remedies with Competition*.¹⁰ PAEs purchase patents from others—individual inventors, universities, failing firms, firms jettisoning poorly performing divisions, firms selling unneeded assets, etc.—and then seek to “assert” those patents either through licensing, or more commonly, through litigation. This activity stands in contrast to entities that develop and transfer technology through patent licensing. It is therefore the PAEs that most closely fit the literature’s descriptions of patent trolls, although some scholars have written in defense of PAEs as well.¹¹

In this article I use the word *troll* when referring to others’ use of the word. I use the term “NPE” when discussing the broad category of all non-practicing entities. And I use the term PAE when discussing the narrower subset of NPEs that do not invent themselves but instead focus on litigating patents acquired from others.

II. THE EMPIRICAL LITERATURE ON “TROLLS”

The empirical literature on patent trolls is relatively new and still underdeveloped. Nonetheless, the existing papers that have attempted to quantify the nature and impact of PAEs and their lawsuits are instructive. I begin with a paper that attempts to profile PAEs. I then move to the relatively larger set of papers that examine PAE-initiated lawsuits.

⁹ In short, I equate troll-like behavior with patent holdup. For a discussion of patent holdup, see FED. TRADE COMM’N, *THE EVOLVING IP MARKETPLACE: ALIGNING PATENT NOTICE AND REMEDIES WITH COMPETITION* (2011) [hereinafter *FTC IP Report*], available at <http://www.ftc.gov/os/2011/03/110307patentreport.pdf>.

¹⁰ The *FTC IP Report* was the culmination of over a year of investigation by the FTC, which included hearings and workshops held around the country. Another commonly used term is “patent aggregator,” denoting an entity that purchases patents originating with others.

¹¹ In particular, such entities could operate as “market makers,” much like their financial counterparts, creating liquidity and smoothing transactions over IP. See, e.g., James F. McDonough III, *The Myth of the Patent Troll: An Alternative View of the Function of Patent Dealers in an Idea Economy*, 56 *EMORY L.J.* 189 (2006). For a balanced look at the pros and cons of trolls, see John Johnson, Gregory K. Leonard, Christine Meyer & Ken Serwin, *Don’t Feed the Trolls?*, 42 *LES NOUVELLES* 487 (2007). While theories on both sides of this debate abound, empirical research is needed to test the various positive and negative assertions that have been made about PAEs and trolls.

Feldman and Ewing present a thorough study of the typically secretive operations of PAEs, whom they dub “mass aggregators.”¹² They report that the first aggregator of noticeable size emerged in the early 1990s: Acacia Research Corporation (Acacia) was founded in 1993 and went public in December 2002. According to the company’s website, “Acacia Research’s subsidiaries partner with inventors and patent owners, license the patents to corporate users, and share the revenue. Acacia controls over 200 patent portfolios” spanning a broad array of technologies.¹³ Feldman and Ewing report that the firm “has executed more than 1,000 license agreements across 104 technology licensing programs.”¹⁴ Despite the numerous license agreements reported by the company on its website, Feldman and Ewing find that Acacia is “among the most litigious of the non-practicing entities.”¹⁵ They relay that, “According to one report, the company and its subsidiaries have been plaintiffs in 280 patent lawsuits and defendants (presumably from declaratory judgment actions) in still more litigations.”¹⁶

According to Feldman and Ewing’s research, the largest PAE is Intellectual Ventures. Founded by ex-Microsoft executive Nathan Myhrvold, Intellectual Ventures is estimated to hold between 30,000 and 60,000 patents worldwide.¹⁷ However, the firm did not begin to litigate patents directly until quite recently. Instead, it relied on a tactic referred to as “privatizing,” taking a reference from 18th century warfare. Using its vast, extremely complicated and opaque network of at least 1,276 holding companies and other affiliated entities, Intellectual Ventures would sell a patent to an aggressive non-practicing private party, who would then be free to—and could be expected to—sue any and all potential infringers, while Intellectual Ventures retains a license to the transferred patents for its subscribing members.

Intellectual Ventures apparently changed strategies in late 2010. In December of that year, the company filed three large patent infringement suits in its own name.¹⁸ It has since filed additional suits in other jurisdictions, including cases before the International Trade Commission. Most recently, it filed suit against AT&T, T-Mobile, and Sprint Nextel in a single lawsuit over fifteen different patents.¹⁹

With respect to the impact of such litigation, Catherine Tucker studies one particular lawsuit filed by Acacia. The suit asserted two medical imag-

¹² Robin Feldman & Thomas Ewing, *The Giants Among Us*, 2012 STAN. TECH. L. REV. 1 (2012).

¹³ See *About Us*, ACACIA RESEARCH CORP., http://acaciaresearch.com/aboutus_main.htm (last visited July 28, 2012).

¹⁴ Feldman & Ewing, *supra* note 12, at 72.

¹⁵ *Id.* at 73.

¹⁶ *Id.*

¹⁷ *Id.* at 24.

¹⁸ *Id.* at 70.

¹⁹ Patrick Anderson, *Intellectual Ventures Flexes Some Patent Muscle*, PATENTLY-O (Feb. 19, 2012, 2:34 PM), <http://www.patentlyo.com/patent/2012/02/index.html>.

ing software patents and named fourteen distinct defendants.²⁰ Tucker finds the litigation had a dramatic short-term impact on the defendants' investments in new software versions: new releases of imaging software fell to zero for the full set of defendants during the litigation period, despite no change in measured demand from hospitals and no drop in textual medical software developed by the same defendant firms.²¹

Bessen, Ford, and Meurer analyze PAE litigation more broadly in their empirical study. They employ a dataset collated by PatentFreedom, "an organization devoted to researching and providing information on NPE behavior and activities," which purports to record litigation instigated by NPEs.²² PatentFreedom's dataset is private, but the firm's website offers its definition of NPE used for its data collection:

We define an NPE as any entity that earns or plans to earn the majority of its revenue from the licensing or enforcement of its intellectual property. While there are other definitions one might consider, our reason for using this one is clear. Because they do not sell products or services (other than the licensing of their patents), NPEs typically do not infringe on the patent rights contained in others' patent portfolios. As a result, they are essentially invulnerable to the threat of counter-assertion, which is otherwise one of the most important defensive—and stabilizing—measures in patent disputes.

For companies facing it, NPE litigation is therefore particularly challenging. It can be highly distracting to management, which must pay money to outside counsel to defend itself, or to the "other side" in order to secure a license, or both. . . .

NPEs are not all cut from the same cloth. In contrast to widely-held perceptions, approximately 60% of NPEs identified by PatentFreedom are asserting patents originally assigned to them, and another 15% are asserting a blend of originally assigned and acquired patents . . .²³

Using this dataset, Bessen, Ford, and Meurer find that patent litigation with an NPE plaintiff looks quite different from patent litigation initiated by other entities. In particular, they report that NPE litigation "is focused on software and related technologies, it targets firms that have already developed technology, and most of these lawsuits involve multiple large compa-

²⁰ Catherine Tucker, *Patent Trolls and Technology Diffusion* 8-10 (Nov. 23, 2011) (unnumbered working paper), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1976593.

²¹ *Id.* at 16-28. Of course, this may simply be a rational cost-benefit response to the risks entailed in willful infringement charges. Professor Tucker has not yet examined whether the effects continue after the conclusion of the litigation.

²² James Bessen et al., *The Private and Social Costs of Patent Trolls* 9 (Bos. Univ. Sch. of Law, Law & Econ. Research Working Paper Grp., Working Paper No. 11-45, 2011), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1930272.

²³ *What is an NPE?*, PATENT FREEDOM, <https://www.patentfreedom.com/background-npe.html> (last visited July 28, 2012).

nies as defendants.”²⁴ This picture is consistent with that presented by Tucker in her study of the Acacia medical imaging software patent lawsuit.

In addition, the authors “find that NPE lawsuits are associated with half a trillion dollars of lost wealth to defendants from 1990 through 2010.”²⁵ They note that the broader literature on litigation establishes that defendant lost wealth—damages, lawyer and expert fees, plus market capitalization losses—is generally a deadweight loss to society, as relatively little money flows to plaintiffs compared to the total loss.²⁶ Moreover, the authors do not find that defendants’ aggregate losses represent transfers to inventors, so they conclude that the “loss of incentives [to innovate] to the defendant firms is not matched by an increase in incentives to other inventors.”²⁷ However, it is important to note that this finding applies only to public NPEs, likely a small subset of all NPEs.

Even though PAE-centric studies and reports in the popular press may give the impression that PAEs are the only entities filing patent lawsuits,²⁸ broader academic studies of litigation find that these entities’ lawsuits comprise a more modest share of patent infringement litigation than the headlines might suggest. For example, Chien finds that NPEs²⁹ filed 17% of high tech patent infringement lawsuits between 2000 and 2008.³⁰ This is a far smaller percentage than the 41% share of patent infringement suits filed by patent-practicing public firms (aka the “giants”).³¹ However, when Chien accounts for the fact that NPEs are more likely to file suit against multiple defendants in a single case, where defendants are linked only by their alleged infringement of the NPE’s patents (such as in the Acacia case study that Tucker presents), the NPE share jumps considerably.³² This adjusted assessment puts NPEs roughly on par with large public firms, but their share is still half as much as practicing entities as a group. As a proportion of distinct defendants, Chien reports that NPEs represent 26% of all high-tech plaintiffs while public practicing firms comprise 30% and private

²⁴ Bessen et al., *supra* note 22, at 2.

²⁵ *Id.*

²⁶ *Id.* at 5 (citing Sanjai Bhagat & Roberta Romano, *Event Studies and the Law: Part I: Technique and Corporate Litigation*, 4 AM. L. AND ECON. REV. 141 (2002)).

²⁷ *Id.* at 2. One caveat to the authors’ conclusion is that it can be difficult to estimate future deterrence effects. For example, the imposition of such costs on infringers may cause future potential infringers to seek licenses.

²⁸ See, e.g., Dan Frommer, *Patent Troll NTP Sues Apple, Google, HTC, Microsoft, AND Motorola Over Email Patents*, BUSINESS INSIDER (July 9, 2010), http://articles.businessinsider.com/2010-07-09/tech/30061819_1_donald-e-stout-tom-campana-patent-claims.

²⁹ She categorizes universities under a separate “non-profit” category, but otherwise does not differentiate between PAEs and upstream specialists. Chien, *supra* note 4, at 1596.

³⁰ *Id.* at 1572.

³¹ *Id.* at 1600.

³² *Id.* at 1601.

practicing firms comprise 35%.³³ Moreover, Chien finds that lawsuits in certain high-tech areas have a disproportionate share of NPE plaintiffs: 39% of infringement cases over computer software or hardware products are filed by NPEs, and NPEs brought 40% of the infringement cases over finance-related patents.³⁴

Chien's finding on the technological concentration of NPE lawsuits is corroborated in a study by Michael Risch.³⁵ Risch finds that most NPE litigated patents fall under the U.S. Patent classifications devoted to communications and computers. Surprisingly, Risch reports that only 8% of the NPE litigated patents in his study fall under USPTO Class 705—the class most closely associated with business method patents. He also finds that chemical and drug patents are rarely enforced by NPEs. Touching on another aspect of the troll debate, Risch states that “traditional patent quality measures [e.g., patent citations] imply at the very least that NPE patents look a lot like other litigated patents.”³⁶ Despite the similarity of observable patent measures, however, Risch notes that “trolls almost never win infringement judgments.”³⁷

Given their potential reach, the massive portfolio holdings of some, and the secrecy with which they appear to operate,³⁸ it is no surprise that patent “trolls” have captured the attention of scholars and policymakers. As is discussed next, however, the patent assertion entities are not the only intimidating creatures in the patent world. Traditional public firms that practice their own patents wield significant heft as well.

III. PORTRAIT OF A “GIANT”

In the prior era—before the emergence of PAEs, before upstream R&D specialization was practical or even feasible in many sectors, and before universities had created their own patent licensing offices³⁹—patents

³³ *Id.* at 1601-02.

³⁴ *Id.* at 1600.

³⁵ Michael Risch, *Patent Troll Myths*, 42 SETON HALL L. REV. 457 (2012).

³⁶ *Id.* at 481; see also Sannu K. Shrestha, Note, *Trolls or Market-makers? An Empirical Analysis of Nonpracticing Entities*, 110 COLUM. L. REV. 114 (2010).

³⁷ Risch, *supra* note 35, at 481. Allison, Lemley, and Walker come to the same conclusion; they report a win rate for NPEs of 9.2%, compared to 10.7% for the average patent holder. John R. Allison, Mark A. Lemley & Joshua Walker, *Patent Quality and Settlement among Repeat Patent Litigations*, 99 GEO. L. J. 677, 680-81 (2011).

³⁸ A few others include the following: RPX (a patent defense fund), Round Rock Research (acquired portfolio from Micron Technology), and Transpacific IP Ltd. (based in Taiwan and with offices throughout Asia). See generally Feldman and Ewing, *supra* note 12.

³⁹ That is, in the late 19th century through much of the 20th century, but before the Bayh-Dole Act, the Stevenson-Wydler Act, the Semiconductor Chip Protection Act, the creation of the Court of Appeals for the Federal Circuit, and before landmark cases, like *State St. Bank & Trust Co. v. Signature Fin.*

were obtained primarily by manufacturing firms. These were the entities that could afford sustained research efforts, the fruits of which were incorporated into the firms' own products with little thought of outside licensing. As significant new business models untethered to manufacturing emerged, many industry observers appear to have lost sight of the important role that the large do-it-all firms continued to play. The fact they still hold considerable clout over IP in the marketplace was, for a time anyway, forgotten.

But as already noted above, firms that practice their patents are not just the targets of litigation by PAEs, they are actually the most common patent litigators, at least in the high-tech space.⁴⁰ This is not surprising, as efforts to license patents or protect products from infringement are often only successful with the threat of enforcement behind them.

Many large manufacturing firms hold massive patent portfolios. Such holdings are useful in a number of respects. First, they aid in establishing détente with other large manufacturing entities. With large IP portfolios on both sides of the bargaining table, cross-licensing is the most common outcome; neither party will want to sue the other for patent infringement unless it is deemed the only route to resolving a dispute. Second, patents on non-core technologies can provide practicing firms with a source of additional revenue via licensing. Licensing non-core patents can also diffuse the company's technologies throughout an industry and—where non-core patents cover inventions that enable or complement core technologies—spur core product sales.⁴¹ On the darker side, however, some worry that such patent arsenals in the hands of large incumbent firms can also be used to wipe out emerging competitive threats from smaller rivals that do not have armies of patent lawyers or the financial resources to amass huge patent portfolios.⁴² Threatening patent litigation against smaller competitors can be enough to dry up the entrant's sources of financing.⁴³

As an illustration of a patent giant, consider IBM, the poster child for an old-school manufacturing firm that learned early on to embrace the new

Grp., Inc., 149 F.3d 1368 (Fed. Cir. 1998) and *Diamond v. Diehr*, 450 U.S. 175 (1981), all of which have led to a more IP-focused marketplace.

⁴⁰ Chien, *supra* note 4, at 1600-01.

⁴¹ As an example of this latter effect, Bluetooth technology is licensed royalty free by relevant patent holders because its increased adoption leads to increased sales of high tech goods. For further discussion, see Anne Layne-Farrar & Josh Lerner, *To Join or Not to Join: Examining Patent Pool Participation and Rent Sharing Rules*, 29 INT'L. J. INDUS. ORG. 294 (2011).

⁴² This tactic is referred to as "predatory litigation." See, e.g., PIERRE REGIBEAU & KATHARINE ROCKETT, ASSESSMENT OF POTENTIAL ANTICOMPETITIVE CONDUCT IN THE FIELD OF INTELLECTUAL PROPERTY RIGHTS AND ASSESSMENT OF THE INTERPLAY BETWEEN COMPETITION POLICY AND IPR PROTECTION (2011), available at http://ec.europa.eu/competition/consultations/2012_technology_transfer/study_ipr_en.pdf.

⁴³ Josh Lerner, *Patenting in the Shadow of Competitors*, 38 J.L. & ECON. 463 (1995).

world order of IP licensing. According to the following on IBM's website:⁴⁴

In 2007, IBM received 3,125 U.S. patents from the USPTO. This is the fifteenth consecutive year that IBM has received more US patents than any other company in the world. In addition to delivering these innovations through its products and services, IBM maintains an active patent and technology licensing program.

With continual large accretions made each year, IBM's current patent portfolio consists of around 40,000 active patents.⁴⁵ In addition to encouraging firms to license those patents, IBM has also been a regular patent infringement plaintiff. Some suits have been highly controversial, such as when IBM accused a French open source software entity of infringing over one hundred of its patents.⁴⁶ In other instances, IBM's litigation fits the traditional mold discussed above. For example, IBM sued Amazon.com in 2006 after nearly four years of attempts to resolve alleged infringement issues through negotiations.⁴⁷

Texas Instruments (TI) is another giant well known for its relatively aggressive stance on IP. Its patent portfolio size is not far behind IBM's, at around 38,000 patents worldwide, and that count has been growing on the order of 1,000-plus patents each year.⁴⁸ And TI has also been on both sides of patent litigation.⁴⁹ For instance, in the 1980s TI sued nine Korean and Japanese chip makers in order to increase their royalty payments.⁵⁰

With such vast patent portfolios listed among their assets, it is no wonder that manufacturing giants' patents frequently make their way onto technology markets. And that is just what happened in 2011, when numerous patent auctions, acquisitions, and rights transfers took place within the wireless sector. The following section describes the flurry of activity occurring in 2011.

⁴⁴ See *Intellectual Property and Licensing*, IBM, <http://www.ibm.com/ibm/licensing/> (last visited Mar. 16, 2012).

⁴⁵ See *Intellectual Property and Licensing-Patents*, IBM, <http://www.ibm.com/ibm/licensing/patents/> (last visited Mar. 16, 2012).

⁴⁶ For further discussion on the controversy stirred by IBM's assertion of over 100 of its patents against a French open source firm, see Eingestellt von Florian Mueller, FOSS PATENTS (Apr. 6, 2010, 8:11 AM), <http://fospatents.blogspot.com/2010/04/ibm-breaks-taboo-and-betrays-its.html>.

⁴⁷ See *Patent Infringement Lawsuits Against Amazon.com*, IBM (Oct. 23, 2006), <http://www.ibm.com/press/us/en/pressrelease/20481.wss>.

⁴⁸ See *TI Fact Sheet*, TEXAS INSTRUMENTS, <http://www.ti.com/corp/docs/company/factsheet.shtml>.

⁴⁹ See Andrew Pollack, *The New High-Tech Battleground*, N.Y. TIMES, July 3, 1988, § 3, at 1.

⁵⁰ Michael Paul Chu, *An Antitrust Solution to the New Wave of Predatory Patent Infringement Litigation*, 33 WM. & MARY L. REV 1341 (1992).

IV. THE “GREAT PATENT BUBBLE OF 2011”⁵¹

It all began with Novell. With its efforts at reinvention failing and its earnings steadily declining, the company began to shop its assets in the spring of 2010. As many as twenty entities registered bids to purchase Novell.⁵² Since most of the registering suitors were private equity firms, there was speculation that a purchase would be primarily motivated, if not solely motivated, by Novell’s extensive patent holdings.⁵³ In the end, a hybrid deal was struck over Novell’s assets. On November 22, 2010, Novell announced that it was being acquired by the software firm Attachmate Corporation.⁵⁴ However, as part of the deal, Novell also made the following announcement regarding its patent portfolio:

Novell also announced it has entered into a definitive agreement for the concurrent sale of certain intellectual property assets to CPTN Holdings LLC, a consortium of technology companies organized by Microsoft Corporation, for \$450 million in cash, which cash payment is reflected in the merger consideration to be paid by Attachmate Corporation.⁵⁵

The “certain intellectual property assets” referenced in the announcement were some 882 patents, described as a “treasure trove” by some knowledgeable observers.⁵⁶

Attachmate’s acquisition of Novell was consummated in April 2011, but competition authorities at the German Federal Cartel Office and the U.S. Department of Justice (DOJ) questioned the patent portfolio acquisition by the CPTN consortium.⁵⁷ Of particular concern was the fact that Novell, once a big player in the Unix operating system and hence important for the open source operating system Linux, had made certain licensing commitments to the open source community. In order to move the deal

⁵¹ Richard Waters, *Patent Hunting is Latest Game on Tech Bubble Circuit*, FIN. TIMES (July 27, 2011), <http://www.ft.com/intl/cms/s/0/16025f76-b868-11e0-b62b-00144feabdc0.html#axzz1gSKjhuMc>.

⁵² Matt Asay, *Novell Auction Could Be Patent Troll Bonanza*, CNET NEWS (May 28, 2010 9:04 A.M. PDT), http://news.cnet.com/8301-13505_3-20006248-16.html. Matt Asay is a former Novell employee.

⁵³ *Id.*

⁵⁴ *Novell Agrees to be Acquired By Attachmate Corporation*, NOVELL (Nov. 22, 2010), <http://www.novell.com/news/press/2010/11/novell-agrees-to-be-acquired-by-attachmate-corporation.html>.

⁵⁵ Note that Novell had already entered into a joint patent agreement with Microsoft, so Microsoft held a license to the patents that were sold in 2011. See *Microsoft and Novell Announce Broad Collaboration on Windows and Linux Interoperability and Support*, MICROSOFT (Nov. 2, 2006), <http://www.microsoft.com/presspass/press/2006/nov06/11-02MSNovellIPR.msp>.

⁵⁶ Asay, *supra* note 52.

⁵⁷ Jon Brodtkin, *Novell Patent Sale to Shield Microsoft, Apple, EMC and Oracle from Lawsuits*, NETWORK WORLD (Apr. 8, 2011), <http://www.networkworld.com/news/2011/040811-novell-patents.html>.

along while the competition agencies wrestled with the cartel aspects and prior licensing obligations, it was decided—in the words of Open Source Initiative President Michael Tiemann—that “CPTN will now only exist for long enough to distribute the shares equally among the participants in the transaction (no more than three months), and thus will not form a new long-term patent troll itself.”⁵⁸ The final resolution issued by the DOJ made two stipulations.⁵⁹ First, Microsoft must sell back the Novell patents it acquired through CPTN and can maintain only a license to those patents, which it had acquired prior to the Novell patent sale. Second, all of the patents acquired by CPTN must be made available royalty free for open source licensing.

The next patent auction offered up Nortel’s even bigger portfolio. Once “the world’s largest supplier of telecom equipment,”⁶⁰ Nortel filed for bankruptcy in 2009 but was unable to successfully restructure itself. Two and a half years later, in June of 2011, Nortel began auctioning off its most valuable remaining assets: its patents.⁶¹ Nortel’s portfolio consisted of nearly 6,000 patents, including many on wireless, data networking, optical, voice, Internet, and semiconductor technologies, with “the most-wanted ones relat[ing] to emerging 4G [wireless] standards such as long-term evolution (LTE).”⁶²

With the \$450 million price tag from the Novell patent sale still fresh in everyone’s mind, the Nortel patent auction spurred much more up-front discussion than the previous Novell announcement. Immediately, big names in the high tech and telecom sectors, including Apple, Ericsson, Google, Intel, Microsoft, and Research in Motion (RIM), expressed their interest in Nortel’s patent portfolio.⁶³ In fact, Nortel delayed the start of its auction by one week to allow it to evaluate additional bids.⁶⁴ Google’s opening bid of \$900 million (twice the Novell price) was selected as the

⁵⁸ *Id.*

⁵⁹ Press Release, U.S. Dept. of Justice, CPTN Holdings LLC and Novell Inc. Change Deal in Order to Address Department of Justice’s Open Source Concerns (Apr. 20, 2011), available at <http://www.justice.gov/opa/pr/2011/April/11-at-491.html>.

⁶⁰ Josh Kelley, *War of Tech Giants Unfolds with Multi-Billion Dollar Patent Auction [Part One]*, PALO ALTO PATCH (Aug. 8, 2011), <http://paloalto.patch.com/articles/war-of-tech-giants-unfolds-with-multi-billion-dollar-patent-auction-part-one> (quoting Andrew Wahl, *The Good, the Bad and the Ugly: Nortel Networks*, CANADIAN BUS. (Mar 30, 2009)), <http://www.canadianbusiness.com/article/14895--the-good-the-bad-and-the-ugly-nortel-networks>).

⁶¹ See Julie Friedman, *With Cleary Presiding, Nortel Patent Auction Could Be Biggest Ever*, THE AMLAW DAILY (June 24, 2011), <http://amlawdaily.typepad.com/amlawdaily/2011/06/nortelpatentauction.html>.

⁶² Dave Goodboy, *Rockstar Wins \$4.5b Nortel Patent Buyout*, BEACON EQUITY RESEARCH (July 12, 2011), <http://www.beaconequity.com/rockstar-wins-4-5b-nortel-patent-buyout-2011-07-12/>.

⁶³ Tom Krazit, *Nortel Delays Mobile Patent Auction One Week as Bidders Get Ready*, MOCO NEWS (June 16, 2011), <http://moconews.net/article/419-nortel-delays-mobile-patent-auction-one-week-as-bidders-get-ready/>.

⁶⁴ *Id.*

“stalking horse,” setting the floor for further bids. Early speculators anticipated that the sale might generate as much as \$1 billion. These guesses proved to be woefully conservative.

According to Reuters, Intel started the four-day auction for the Nortel patents with a \$1.5 billion opening offer, a 67% premium to the base price set by Google.⁶⁵ The bidding went for nineteen rounds, concluding in early July 2011. Reminiscent of the Novell patent purchase, a consortium provided the winning Nortel bid. Rockstar Bidco, a coalition of Apple, Microsoft, EMC, Ericsson, RIM, and Sony, won the patent portfolio with a \$4.5 billion buyout bid.⁶⁶

Finding concerns of collusion similar to those that prompted competition agency review of the Novell deal, the American Antitrust Institute wrote to the Assistant Attorney General for Antitrust, calling for an in-depth DOJ investigation into Rockstar’s purchase.⁶⁷ It is important, though, to keep in mind the difference in circumstances between the two auctions. Both auctions offered failing firm patent assets for sale, but, unlike Novell, Nortel had no open source licensing encumbrances attached to its portfolio.

The DOJ’s decision, issued on February 13, 2012,⁶⁸ allowed both consortium deals to go through, relying in particular on commitments made by Apple and Microsoft. Specifically, the two companies committed to license any of the patents relevant for industry standards on fair, reasonable, and non-discriminatory (FRAND) terms and not to seek “offensive” injunctive relief—that is, to only seek injunctive relief after an opposing party sought it first. The decision indicates that joint bidding in patent auctions, at least with some assurances of reasonable licensing, is a viable strategy going forward.

The criticisms over consortium purchases of the patents were not confined to antitrust allegations of collusion. Google swiftly cried foul over both the Novell and Nortel auctions, coining the nickname “Micropple” for the Microsoft–Apple led coalitions (first CPTN with Novell, then Rockstar with Nortel). Google complained that these firms (and their cohorts) were “engaged in a ‘hostile’ patent war against the search giant.”⁶⁹ David

⁶⁵ Kelley, *supra* note 60.

⁶⁶ Goodboy, *supra* note 62.

⁶⁷ American Antitrust Institute Warns of Anticompetitive Effects from Wireless Technology Consortium’s \$4.5 Billion Purchase of Nortel’s Entire Patent Portfolio, AM. ENTERPRISE INST. (July 6, 2011), <http://www.antitrustinstitute.org/content/american-antitrust-institute-warns-anticompetitive-effects-wireless-technology-consortiums-4>.

⁶⁸ Press Release, U.S. Dept. of Justice, Statement of the Department of Justice’s Antitrust Division on Its Decision to Close Its Investigations of Google Inc.’s Acquisition of Motorola Mobility Holdings Inc. and the Acquisitions of Certain Patents by Apple Inc., Microsoft Corp. and Research in Motion Ltd. (Feb. 13, 2012), *available at* <http://www.justice.gov/opa/pr/2012/February/12-at-210.html>.

⁶⁹ David Drummond, *When Patents Attack Android*, GOOGLE BLOG (Aug. 3, 2011), <http://googleblog.blogspot.com/2011/08/when-patents-attack-android.html>.

Drummond, Google's Chief Legal Officer, argued the following on the official Google blog that:

They're doing this by banding together to acquire Novell's old patents (the "CPTN" group including Microsoft and Apple) and Nortel's old patents (the "Rockstar" group including Microsoft and Apple), to make sure Google didn't get them; seeking \$15 licensing fees for every Android [mobile operating system] device; attempting to make it more expensive for phone manufacturers to license Android (which we provide free of charge) than Windows Phone 7; and even suing Barnes & Noble, HTC, Motorola, and Samsung. Patents were meant to encourage innovation, but lately they are being used as a weapon to stop it.⁷⁰

Interestingly, Google had been invited to join the Novell bidding effort with CPTN, but declined to do so.⁷¹ Google pointed out that a "joint acquisition of the Novell patents that gave all parties a license would have eliminated any protection these patents could offer to Android against attacks from Microsoft and its bidding partners."⁷² Of course, a license to the patents for Google, Apple and Microsoft would have prevented any patent "attacks" involving these patents, assuming patent exhaustion or some other pass-through of rights to Android-reliant device makers. However, by sharing the license Google would not have obtained an IP club of its own to use against other patent assertions. Thus, one observer opined that

once Google figured out that they wouldn't be the only ones with access to these patents, and that it would basically give them a stalemate, allowing them no leverage over patents gained in [other patent auctions], or elsewhere, it effectively dropped its bid. If it couldn't gain some sort of decisive advantage with the purchase, then it figured it was a waste of money.⁷³

It is certainly not surprising that Google would want to amass a patent portfolio of its own. As a "new entrant" among several established large mobile telecom firms, Google is currently more of a wireless telecom "dwarf"—it is one of the few firms in the industry without a sizeable patent portfolio. As of early 2011 fewer than 2,000 U.S. patents were under Google's control,⁷⁴ including acquisitions of mobile phone-related patents

⁷⁰ *Id.*

⁷¹ Matthew Panzarino, *Google Says It Didn't Want Novell Patents If Everyone Got Them [Updated]*, THENEXTWEB (Aug. 4, 2011), <http://thenextweb.com/google/2011/08/04/google-says-it-didnt-want-novell-patents-if-everyone-got-them/>.

⁷² Drummond, *supra* note 69.

⁷³ Panzarino, *supra* note 71.

⁷⁴ Bill Slawski, *Google Patents Updated*, SEO BY THE SEA (Feb. 6, 2011, 3:30 AM), <http://www.seobythesea.com/2011/02/google-patents-updated/>. A search of the USPTO database conducted December 14, 2011, found 956 patents assigned to Google, but this count misses the recent reassignments resulting from Google's purchases, as discussed *infra* note 78 and accompanying text and *infra* note 83 and accompanying text.

that Google had received from the Myriad Group and from Verizon.⁷⁵ As discussed above, in relation to patent assertion entities, having a portfolio of your own is an important defensive mechanism for patent-practicing firms—hence Google’s interest in Novell’s and Nortel’s patent portfolios and its disappointment in not acquiring them.

Despite Google’s rhetoric following the Nortel auction about “bogus” patents stopping genuine innovation,⁷⁶ the company has not let its patent auction bidding frustrations deter it from seeking other sources of patent acquisition. Indeed, in late July 2011, Google concluded a purchase deal with IBM for 1,030 patents thought to be relevant for the Android mobile operating system.⁷⁷

That same month, July of 2011, InterDigital, yet another wireless telecom company, announced that it too “was looking at putting itself up for sale: with a market value of \$3.2 billion even before any auction begins.”⁷⁸ When the *Wall Street Journal* announced that Google was considering making a bid, InterDigital’s share price increased by 15%.⁷⁹ No deal had been consummated by the end of 2011 however and Google shifted its attention elsewhere.⁸⁰ Finally, in June of 2012, InterDigital sold around of its 1,700 wireless technology patents to Intel, for \$375 million.⁸¹

Google’s new object of affection turned out to be Motorola Mobility Holdings Inc. (MMI)—Motorola’s subsidiary that makes Android smart phones and tablets, among other things. In August of 2011, Google announced its purchase of MMI for \$12.5 billion.⁸² With the acquisition approved by the competition authorities in February 2012,⁸³ Google gained

⁷⁵ Bill Slawski, *Is Google Now a Phone Company?*, SEO BY THE SEA (Dec. 21, 2010, 4:26 PM), <http://www.seobythesea.com/2010/12/is-google-now-a-phone-company/>.

⁷⁶ Drummond, *supra* note 69.

⁷⁷ Jolie O’Dell, *Google Buys 1,030 IBM Patents, Girding Its Loins for Android Lawsuit*, VENTURE BEAT (July 29, 2011), <http://venturebeat.com/2011/07/29/google-ibm-patents/>.

⁷⁸ Waters, *supra* note 51.

⁷⁹ Shira Ovide, *Meet Google’s Latest Takeover Target: InterDigital*, WALL ST. J. BLOG (July 20, 2011, 9:41 AM), <http://blogs.wsj.com/deals/2011/07/20/meet-googles-latest-takeover-target-interdigital/>.

⁸⁰ Due to lack of interest, InterDigital withdrew its sale offer in late January 2012. See Michael J. De La Merced, *InterDigital Calls Off Patent Sale*, N.Y. TIMES (Jan. 23, 2012), <http://dealbook.nytimes.com/2012/01/23/interdigital-said-to-call-off-patent-sale/>.

⁸¹ Sinead Carew, *Intel to Buy Inter Digital Patents for \$375 Million*, REUTERS (June 18, 2012, 1:59 PM), <http://www.reuters.com/article/2012/06/18/us-interdigital-intel-idUSBRE85H17S20120618> (“InterDigital Inc. said on Monday it had agreed to sell to Intel Corp about 1,700 wireless technology patents for \$375 million, sending InterDigital shares up 27 percent.”).

⁸² Sayantani Ghosh, *InterDigital Skids After Google Goes for Motorola Mobility*, REUTERS (Aug. 15, 2011, 12:10 PM), <http://www.reuters.com/article/2011/08/15/us-interdigital-shares-idUSTRE77E3FA20110815>.

⁸³ The European Commission delayed the acquisition on December 12, 2011, citing the need for additional documents. See James Kanter, *Google Acquisition of Motorola Delayed in Europe*, N.Y. TIMES, (Dec. 12, 2011), <http://www.nytimes.com/2011/12/13/technology/google-acquisition-of->

control of MMI's 17,000 granted patents, plus its 7,500 pending patent applications,⁸⁴ putting Google one step closer to achieving "giant" status in the mobile telecom world, adding to its already leading position in trade secret-reliant internet search.

Not to be left out, Nokia joined the "patent bubble" in the fall of 2011. On September 1, 2011, Mosaid, a firm that "licenses patented intellectual property in the areas of semiconductors and communications technologies, and develops semiconductor memory technology,"⁸⁵ announced that it had acquired 1,200 Nokia standards-essential wireless patents and 800 wireless implementation patents.⁸⁶

The details of the Nokia deal are different, and more convoluted, than the earlier 2011 mobile patent acquisitions. In particular, while the patents were all filed by Nokia, they are held by Core Wireless Licensing, a Luxembourg company. Under the terms of the deal, Core Wireless became a wholly-owned subsidiary of Mosaid but did not transfer the patents to Mosaid. Moreover, Mosaid did not pay directly for the Core Wireless acquisition, but instead announced that it "will fund its acquisition of the portfolio through royalties from future licensing and enforcement revenues."⁸⁷ Mosaid, through Core Wireless, will receive only one-third of any "licensing and enforcement revenues," however.⁸⁸ Nokia and—in an interesting twist—Microsoft will share the remaining revenue.⁸⁹ Microsoft entered this deal, which it defines as a "passive economic interest,"⁹⁰ through the wide-reaching Windows Phone collaboration deal it struck with Nokia in

motorola-delayed-in-europe.html. On February 13, 2012—the same day the DOJ approved the Novell and Nortel acquisitions—the EC approved Google's MMI deal, albeit somewhat reluctantly, stating that it would keep a close eye on the company. See Diane Bartz & Foo Yun Chee, *Google Deal Gets U.S., EU Nod to Buy Motorola Mobility*, REUTERS (Feb. 14, 2012, 2:39 AM), <http://www.reuters.com/article/2012/02/14/us-google-motorola-eu-idUSTRE81C1HE20120214>.

⁸⁴ "Google fits the mold of a company at which revenue growth has outpaced its ability to generate its own patents and therefore has been forced to buy aggressively until its internal efforts catch up." Mike McLean, *Google and Motorola—A Match Made in Patent Heaven?*, ELECTRONICS DESIGN NETWORK (Sep. 15, 2011), http://www.edn.com/article/519354-Google_and_Motorola_a_match_made_in_patent_heaven_.php.

⁸⁵ Press Release, Mosaid, Mosaid Acquires 1,200 Nokia Standards-Essential Wireless Patents and 800 Wireless Implementation Patents (Sep. 1, 2011), available at <http://www.mosaid.com/corporate/news-events/releases-2011/110901.php>.

⁸⁶ *Id.*

⁸⁷ *Id.*

⁸⁸ Chris Velazco, *Mosaid Acquires 2,000+ Nokia Patents, Will Handle Licensing & Litigation For A Cut*, TECHCRUNCH (Sep. 1, 2011), <http://techcrunch.com/2011/09/01/mosaid-acquires-2000-nokia-patents-will-handle-licensing-litigation-for-a-cut/>.

⁸⁹ *Id.*

⁹⁰ Mary Jo Foley, *Microsoft Weighs in on Mosaid-Nokia Patent Deal*, ZDNET (Sep. 2, 2011), <http://www.zdnet.com/blog/microsoft/microsoft-weighs-in-on-mosaid-nokia-patent-deal/10523>.

the spring of 2011.⁹¹ In this arrangement, Mosaid effectively operates as a type of collections agent on behalf of Nokia and Microsoft, taking a one-third cut of any licensing or litigation proceeds.

This last of the many wireless telecom patent deals of 2011 could have far-reaching implications. According to Mosaid, 1,200 of the Nokia patents and applications “have been declared essential to second, third and fourth-generation communications standards, including GSM (Global Systems for Mobile communications), UMTS/WCDMA (Universal Mobile Telecommunications Service/Wide-Band Code Division Multiple Access) and LTE (Long Term Evolution).” Hinting at extensive efforts to license or litigate this portfolio, or both, Mosaid’s press release went on to state that, “Based on its extensive experience in the industry, MOSAID believes that revenues from licensing, enforcing and monetizing this wireless portfolio will surpass the Company’s total revenues since its formation in 1975.”⁹² In 2011, Mosaid reported revenues of \$80 million (Canadian dollars), so it is expecting a substantial return indeed on its purchase.

As for Mosaid’s choice between licensing and litigation, litigation seems the more likely route. Mosaid has a history of filing patent infringement lawsuits, to name a few, suits against ASUSteK, Asus Computer, Canon, Dell, Huawei Technologies, HTC, Intel, Lexmark, RIM, Sony Ericsson, and Wistron.⁹³ One industry analyst said the following:

It would be in Mosaid’s best interest to play the bulldog and aggressively pursue not only licensing opportunities, but hefty settlements against companies that infringe on the Nokia patents. Meanwhile, Nokia benefits from whatever Mosaid manages to bring in, but without looking like they’re going on a wild suing spree.⁹⁴

Having established some background information on the topic, this paper now turns to the mobile-patent-related litigation taking place in 2011.

V. THE MOBILE LEGAL BATTLEFIELD

Why spend such significant resources on acquiring patent portfolios through auctions and other transfers? To enforce them, of course. Table 1 below summarizes the lawsuits with activity in 2011 or 2012 focusing solely on mobile telecom-related lawsuits, in line with the patent auctions and

⁹¹ Mary Jo Foley, *Microsoft and Nokia Finalize Their Windows Phone Collaboration Agreement*, ZDNET (Apr. 21, 2011), <http://www.zdnet.com/blog/microsoft/microsoft-and-nokia-finalize-their-windows-phone-collaboration-agreement/9255?tag=content;siu-container>.

⁹² Mosaid Press Release, *supra* note 85.

⁹³ Foley, *supra* note 90.

⁹⁴ Velazco, *supra* note 93.

acquisitions discussed above, and limited to cases with activity in 2011 or the first quarter of 2012. Even with these restrictions, the list is extensive.

Table 1. Patent Infringement Litigation in Telecom, Cases Active in 2011, in order of filing⁹⁵

Plaintiff	Defendant	Date Filed	Court & Case No.	Notes
InterDigital	Nokia	September 2007	ITC	Nokia accused of infringing 3G patents
Nokia	InterDigital	February 2008	Southern District of New York	Countersuit; District Court dismissed Nokia's suit; 2nd Circuit ultimately found for InterDigital
IPCom	HTC	April 2008	German court	HTC accused of infringing IPCom's 2G and 3G patents
HTC	IPCom	November 2008	District of District of Columbia, 08-CV-1897	Request for declaratory judgment that HTC does not infringe IPCom patents on 2G and 3G wireless standards
IPCom	T-Mobile	November 2008	German court	Phone maker accused of infringing IPCom's 2G and 3G patents

⁹⁵ Table compiled by author.

Plaintiff	Defendant	Date Filed	Court Case No.	& Notes
Nokia	Apple	October 2009	District of Delaware, 1:09-cv-00791-UNA	Apple accused of infringing Nokia patents on WiFi, 2G and 3G mobile; settled June 2011
Apple	Nokia	December 2009	District of Delaware, C.A. 09-791-GMS	Countersuit claiming non-essentiality, non-infringement, invalidity of Nokia's patents, alleging Nokia infringes Apple patents; settled in June 2011
Kodak	Apple, RIM	February 2010	ITC, 337-TA-703	iPhone and Blackberry accused of infringing one Kodak patent
Apple	HTC	March 2010	District of Delaware, 1:99-mc-09999; ITC 337-TA-710	ITC ruled in Apple's favor on 2 patents out of 10 covering smart-phones

Plaintiff	Defendant	Date Filed	Court & Case No.	Notes
Apple	Kodak	April 2010	ITC, 337-717	Countersuit alleging Kodak is infringing two Apple patents; ITC ruled in Kodak's favor (non-infringement) in May 2011
Oracle	Google	August 2010	Northern District of California, No. C 10-03561 WHA	Google Android alleged to infringe Oracle patents and copyright.
Motorola	Apple	October 2010	District of Delaware, 1:10-cv-00867-UNA	Apple iPhones, iPads, iTouches, and some Macs accused of infringing
Apple	Motorola	October 2010	Western District of Wisconsin, 10-CV-662	Countersuit alleging Motorola Android phones accused of infringing 6 of Apple's phone patents
Microsoft	Motorola	October 2010	Western District of Washington, C10-01577-RSM; ITC 337-TA-744	Motorola Android phones accused of patent infringement

Plaintiff	Defendant	Date Filed	Court & Case No.	Notes
Motorola	Microsoft	November 2010	Southern District of Florida, 1:10-CIV-24063; Western District of Wisconsin, WI 10-cv-00699	Countersuit alleging Microsoft Windows Mobile of patent infringement
VirnetX	Siemens and Mitel	January 2011	Eastern District of Texas 6:11-cv-00018-LED	Defendants internet protocol phones and communications devices accused of patent infringement
Microsoft	Barnes & Noble, FoxConn, Inventec	March 2011	Western District of Washington, 2:11-cv-00343; ITC 337-TA-769	After a year of negotiations, Microsoft sued B&N's Nook and its manufacturers FoxConn and Inventec for infringing Android patents
Apple	Samsung	April 2011	Northern District of California, 11-CV-01846-LHK plus cases filed in 13 other countries	Samsung's Galaxy phones and computer tablets alleged to infringe Apple's trade dress, trademarks, and utility and design patents

Plaintiff	Defendant	Date Filed	Court & Case No.	Notes
Samsung	Apple	April 2011	Courts in South Korea, Japan, and Germany	Countersuits alleging Apple infringes Samsung's mobile standards patents
Ericsson	ZTE	April 2011	Courts in the UK, Italy, and Germany	After protracted licensing negotiations failed, Ericsson sued ZTE over its patents
ZTE	Ericsson	April 2011	Chinese State IP Office (SIPO)	Countersuit alleging Ericsson infringes ZTE patents on 2G and 4G mobile standards
Huawei	ZTE	April 2011	Courts in Germany, France, and Hungary	ZTE phones accused of infringing Huawei trademarks and patents on data cards and the 4G mobile standard
ZTE	Huawei	April 2011	SIPO	Countersuit alleging Huawei's 4G patents

Plaintiff	Defendant	Date Filed	Court Case No.	& Notes
Apple	HTC	June 2012	ITC	A follow on to the 2010 case; Apple alleges that HTC's work-arounds were insufficient to avoid infringement
InterDigital	Huawei, Nokia, ZTE	July 2011	District of Delaware 1:2011cv0065 4	Patent infringement suit on 3G phones, USB sticks, mobile hotspots and tablets
VirnetX	Apple	November 2011	Eastern District of Texas, 6:2011cv0056 3	Apple iPads, iPods, and iPhones alleged to infringe VirnetX patents
Intellectual Ventures	AT&T, T-Mobile, and Sprint Nextel	February 2012	District of Delaware, 1:2012cv0019 3	Patent infringement on wireless services

The first striking element of Table 1 is its sheer size. Twenty-seven lawsuits are listed. Even if we consolidate countersuits with the original suit, there are still at least eighteen patent-related lawsuits involving mobile telecomm that were active in 2011 through Q1 of 2012.

The second striking element is the prevalence of the relative newcomers to mobile telecom—the dwarves. Apple is named in ten of twenty-five rows above, Google is named in another four through “proxy fights”

against the Android mobile operating system,⁹⁶ and Microsoft is named in another two.

I argue in the next section that the patent dwarves entering the mobile space are the key reason for both the patent bubble and the extensive litigation currently underway.

VI. THE FORCES BEHIND THE WIRELESS PATENT ACQUISITION AND PATENT LITIGATION BUBBLE

The unprecedented patent acquisition activity taking place in just one year begs the question of what lies behind the “Great Patent Bubble of 2011.” As noted earlier, each of the auctions or acquisitions discussed above was related in some way to wireless phones. The likely catalyst of the patent bubble therefore appears to be the changing wireless competitive landscape.

The three new mobile entrants that emerged in the new millennium created a disruption to the ecosystem equilibrium. Microsoft entered the smart phone arena in 2002 with its Windows Mobile operating system,⁹⁷ although this platform appears to have had only a minor impact on the industry thus far.⁹⁸ In contrast, Apple entered the wireless telecom industry in 2007 with its game-changing iPhone and has steadily amassed share ever since.⁹⁹ Google then introduced the Android wireless operating system in 2008 with multiple manufacturers building to incorporate it, and it too, has rapidly earned share.¹⁰⁰

These three firms are not fragile start-ups—they are established entities, well-funded, and well-positioned to compete strongly in the wireless marketplace. What they all lacked in 2011 however, were significant patent portfolios specific to the wireless space. When seen in this light, Apple’s, Microsoft’s, and Google’s scramble to acquire patents, and each entity’s efforts to prevent the other from doing so, is economically understandable even with the multi-billion dollar portfolio price tags. Once the new en-

⁹⁶ Google asserts that the Android OS is open source software, and thus should not be subject to asserted patents. The industry has clearly challenged that position. See *Welcome to Android*, OPEN SOURCE PROJECT, <http://source.android.com/> (last visited July 26, 2012).

⁹⁷ Chris Tilley, *The History of Windows CE*, HPC FACTOR (Feb. 18, 2001), <http://www.hpcfactor.com/support/windowsce/>.

⁹⁸ Brian X. Chen, *Nokia’s Windows Phones Get a Good Start in Europe*, BITS, N.Y. TIMES (Feb. 21, 2012, 12:42 PM), <http://bits.blogs.nytimes.com/2012/02/21/nokia-lumia-europe/>.

⁹⁹ Sarah Radwanick, *5 Years Later: A Look Back at the Rise of the iPhone*, COMSCORE VOICES (June 29, 2012), http://blog.comscore.com/2012/06/5_years_later_a_look_back_at_the_rise_of_the_iphone.html.

¹⁰⁰ See Brad Cook, *Google Overtakes Apple in U.S. Smartphone Market Share*, MACOBSERVER (Jan. 7, 2011, 9:02 AM), http://www.macobserver.com/tmo/article/google_overtakes_apple_in_u.s._smartphone_market_share/.

trants' desire to acquire wireless-relevant patents became known through the price garnered by the Novell auction, other entities with more wireless patents than cash were on notice and the bubble began to expand.

Wireless ecosystem sales opportunities are huge. Globally, in the fiscal year running from April 2009 to March 2010, wireless network operators earned in the aggregate over \$300 billion in revenues with four operators earning \$50 billion each.¹⁰¹ In 2011, wireless device sales generated \$61.5 billion in revenues.¹⁰² But network revenues and handset sales are not all that is at stake—important ancillary opportunities for the wireless platform, such as mobile search, mobile payments, online advertising, and app sales, are key as well. It is only natural then, that the competitive battle would have spilled over from the product marketplace to the technology licensing marketplace and to the courthouse.

Thus, the spate of Apple, Google, and Microsoft-related patent lawsuits fits within the “market disruption” picture as well. The patent auction and acquisition activity make it clear that eventually, all three entrants will have significant wireless-specific patent arsenals of their own. The early years (2011–2012), however, are when these dwarves are the most vulnerable to patent litigation from each other and from the industry giants—the more established wireless players who already have a large portfolio of patents essential for the wireless telecom standards and who compete directly with the Apple, Android, and Windows wireless platforms in the downstream marketplace.

Seen in this light, one way to view Table 1 is from a competition economics standpoint. As established in the theoretical economics literature, vertically-integrated firms can have incentives to raise their downstream rivals' costs; litigation over upstream patent holdings can be one route to do so.¹⁰³ Viewing the table through an IP lens, however, simply indicates that newcomers to an IP-heavy industry cannot avoid clashes with incumbents—patent licensing (or litigation when licenses are not easy to negotiate) appears to be part of the industry entry fee.

This being said, there seems to be little reason to expect long-lasting patent auction or patent litigation activity related to mobile telecom. If the current dust-up is indeed simply a marketplace reaction to new entrants then the ecosystem will eventually settle down—at least until the next disruptive event occurs.

¹⁰¹ BP Tiwari, *Global Wireless Data Update, Quarterly Wrap-up: Q1,2010* 6, BEYOND4G (Mar. 31, 2010), <http://www.beyond4g.org/wp-content/uploads/2010/08/Mobile-Data-Wrapup-Q1-20101.pdf>.

¹⁰² *Wireless Platforms Market - Global Forecasts and Analysis (2011 - 2016)*, MARKETANDMARKETS (Jan. 2012), <http://www.marketsandmarkets.com/Market-Reports/wireless-platforms-market-535.html>.

¹⁰³ See Anne Layne-Farrar & Klaus M. Schmidt, *Licensing Complementary Patents: “Patent Trolls,” Market Structure, and “Excessive” Royalties*, 25 BERKELEY TECH. L.J. 1121 (2010); see also Klaus M. Schmidt, *Complementary Patents and Market Structure* (CEPR Discussion Paper No. DP7005, Oct. 2008), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1311129.

CONCLUSION

Looking beyond mobile telecom, what broader lessons can we learn from the exciting events of 2011? It seems clear that the emergence of PAEs has forever changed the way everyone—whether a practicing entity or not—views patents. The assertion and the practice of patents appear to have been irrevocably split. This distinction was intentional in the creation of U.S. patent laws¹⁰⁴ but the levels of separation seen today appear unprecedented.

This fact comes with a number of consequences. Among them is a considerably increased notion of patent value—it is hard to imagine a patent auction yielding multi-billion dollar price tags in the era before PAEs came to fame. Indeed, absent the increased separation of patents and their practice, new wireless entrants may not have been able to enter the industry at all without first amassing patent portfolios of their own. This requirement would have increased the cost of market entry considerably and would surely have delayed entry to the detriment of consumers.

Following from market entry considerations, fungible and enforceable patents in the hands of non-practicing entities have already had a positive impact on market structure and competition in several areas of the economy like semiconductors, as noted above.¹⁰⁵ These effects will likely continue to unfold.

Another consequence appears to be increased litigation—it is doubtful that unenforceable patents have much value regardless of the innovative content of the underlying patented technology. That means litigation is an important tool for all patent holders: giant against giant, with a focus on preventing direct rivals from copying key technologies and thereby stealing sales; giant against dwarf, with the goal of restricting entry or softening competition; or NPE against practitioner, with a focus on establishing reasonable royalties for the authorized use of patented technology.

Moreover, as the above list of possible patent litigation motives makes clear, patents can be forces for good (reduced entry barriers, increased innovation, prevention of free riding, etc.) or evil (holdup, raising rivals' costs, market foreclosure, etc.). What matters for a good versus evil determination is not who holds the patents, but rather how those patents are used. NPEs and practicing entities alike can practice anticompetitive patent assertion strategies. As a result, the key, in my view, is whether holdup is profitable and possible for the patent holder whichever business model it might follow.

¹⁰⁴ As patent law makes clear, “THE INVENTOR IS NOT REQUIRED TO REDUCE THE INVENTION TO PRACTICE.” MANUAL OF PATENT EXAMINING PROCEDURE § 2137.01 (8th ed. 2008).

¹⁰⁵ See Geradin et al. *supra* note 5.

It also seems clear that we should not focus too heavily on patent litigation itself, as opposed to patent holdup. As noted above, patent rights are essentially meaningless unless they can be enforced. This is why we see trolls, giants, and dwarves—plus other business models not identified here—engaged in patent litigation. Patent litigation may indeed have social costs, as Bessen, Ford, and Meurer argue, but it is difficult to see how we might reduce those costs without killing the many benefits associated with enforceable patent rights.

In the wake of the 2011 patent reform legislation—the America Invents Act of 2011—however, PAE litigation practices may be dramatically altered.¹⁰⁶ The Act prevents plaintiffs from naming multiple unrelated defendants to a single patent infringement case.¹⁰⁷ This will raise the cost of pursuing numerous defendants because each defendant will require a separate lawsuit. This change could also curb the damages awarded to PAE plaintiffs in any one suit because it affects patent holder leverage and the ability to play one defendant off of another. As a result, the Act could (eventually) reduce PAE patent infringement litigation activities, especially in light of the empirical studies that find NPEs have lower odds of winning patent infringement cases.¹⁰⁸

As for the giants, they appear to be increasingly embracing the full range of possibilities embodied in their patent portfolios—beyond patents' use in cross-licensing other practicing entities. It is unlikely that this genie will be put back in the bottle.

Finally, which consequences endure and which will fade also depends, at least to some degree, on the extent to which competition agencies see fit to intervene. As noted above, the U.S. Department of Justice and the European Union Commission both closely examined the two joint patent acquisitions and Google's purchase of MMI, although all three transactions were ultimately cleared. For many years now, competition agencies around the globe have also taken an active interest in IP that relates to cooperative industry standard setting.¹⁰⁹ We have yet to see agency interest in patent trolls outside of industry standards but it is not unthinkable. With the green light given to joint patent auction bidding, we might see proposals for joint patent licensing where non-practicing patent holders join together for the licensing of their patents. This would not be in a patent pool (which would surely run afoul of competition authorities given the high likelihood of substitute technologies), but rather to consolidate the transaction costs of licensing and enforcing patents analogous to university technology transfer

¹⁰⁶ Leahy-Smith America Invents Act, Pub. L. No. 112–29, 125 Stat. 284 (2011).

¹⁰⁷ *Id.*

¹⁰⁸ See Allison, Lemley & Walker, *supra* note 37.

¹⁰⁹ For the latest example, see the current investigation of Samsung by the European Commission. Press Release, European Commission, Antitrust: Commission Opens Proceedings Against Samsung (Jan. 31, 2012), available at <http://europa.eu/rapid/pressReleasesAction.do?reference=IP/12/89>.

offices or the music collection societies like the American Society of Composers, Authors, and Publishers (ASCAP).

Given the complexity of the effects of NPEs, PAEs, and tradable patent rights on the economy, at least one thing does seem certain: the debate over patent trolls will continue to rage. Before anything conclusive can be said however, more empirical analysis is needed to bring clarity to the debate. We need to understand the impact of troll behavior not just on their lawsuit targets and on public NPEs, but more broadly on individual inventors, small private firms and start-ups (collectively the “hobbits” perhaps?), as well as large practicing entities (both dwarves and giants). We have seen hints that each of these entities has responded to the PAE business model with changes in their own practices, so each of these parties needs to be better understood before we can claim to have a complete picture of the *Brothers Grimm* book of business models.



THE PRIVATE COSTS OF PATENT LITIGATION

*James Bessen & Michael J. Meurer**

ABSTRACT

This paper estimates the total cost of patent litigation. We use a large sample of stock market event studies around the date of lawsuit filings for U.S. public firms from 1984 to 1999. Even though most lawsuits settle, we find that the total costs of lawsuits are large compared to estimated legal fees, estimates of patent value, and research and development spending. By the late 1990s, alleged infringers bore expected costs of over \$16 billion per year. These estimates support the view that infringement risk should be a major concern of policy.

INTRODUCTION

Like any regulatory mechanism, the patent system has benefits and costs, both private and social. Yet little empirical evidence exists about the magnitude of some of these costs, leaving policy analysts to sometimes rely on guesswork. For example, recent policy analysis of patent opposition proceedings in the U.S. has been based on rough estimates of the costs of patent litigation and the social costs of inappropriately-granted patents.¹

In contrast, significant literature estimates the benefits of the patent system, especially private benefits in the form of estimates of patent value² or of the patent premium.³ However, without comparable estimates of pri-

* Research on Innovation and Boston University School of Law, and Boston University School of Law, respectively. Thanks for comments to Megan MacGarvie, Jesse Giummo, Tom Hazlett, John Turner and conference participants at the IIOC, CELS, and The Digital Inventor at George Mason University, and seminars at Harvard, Stanford, the NBER, and the NBER Summer Institute. Thanks also to research assistance from Debbie Koker and Dan Wolf. Contact: jbessen@bu.edu.

¹ See Jonathan Levin & Richard Levin, *Patent Oppositions 2-4* (Stanford Inst. for Econ. Pol'y Research, Discussion Paper No. 01-29, 2002), available at <http://www.ssrn.com/abstract=351900>; see also Bronwyn H. Hall et al., *Prospects for Improving U.S. Patent Quality via Post-Grant Opposition* 8 (Nat'l Bureau of Econ. Research, Working Paper No. 9731, 2004), available at <http://www.nber.org/papers/w9731>.

² See JAMES BESSEN & MICHAEL MEURER, *PATENT FAILURE: HOW JUDGES, BUREAUCRATS, AND LAWYERS PUT INNOVATORS AT RISK* (2008); Ariel Pakes & Mark Schankerman, *The Rate of Obsolescence of Patents, Research Gestation Lags, and the Private Rate of Return to Research Resources*, in *R&D, PATENTS & PRODUCTIVITY* 73 (1984).

³ See Ashish Arora et al., *R&D and the Patent Premium* 43 (Nat'l Bureau of Econ. Research, Working Paper No. 9431, 2005), available at <http://www.nber.org/papers/w9431>.

vate and social costs, it is difficult to conduct either analyses of specific policy changes or a normative analysis of the patent system in comparison to other means of encouraging innovation. For example, Schankerman suggests that the ratio of aggregate patent value to research and development (R&D) constitutes an upper bound measure of the subsidy that patents provide to R&D.⁴ He asserts that this ratio can be used to compare patents to other forms of appropriating returns on invention. But surely this is only an estimate of a *gross* subsidy against which private costs of patents need to be netted out.

This paper takes a step toward quantifying costs by estimating the private costs of patent litigation. Using event study methodology to analyze patent lawsuit *filing*, we find the expected joint loss to the litigating parties is large, and probably much larger than the expected attorneys' fees. This result is a bit surprising because most patent lawsuits settle short of trial, and thus it might seem that average patent litigation costs would not be large.

But attorneys' fees and the indirect costs of litigation can be high even when a patent lawsuit settles before trial. Indirect business costs of patent litigation take many forms. For example, the time managers and researchers spend producing documents, testifying in depositions, strategizing with lawyers, and appearing in court can each disrupt business. Litigation strains the relationship between the parties and may jeopardize cooperative development of the patented technology or cooperation on some other front. In addition, firms in a weak financial position might see their credit costs soar because of possible bankruptcy risk created by patent litigation.

Alleged infringers face additional costs. Preliminary injunctions can shut down production and sales during pending litigation. But even without a preliminary injunction, customers may stop buying an alleged infringer's product. Frequently, products require customers to make complementary investments; customers may not be willing to make these investments if a lawsuit poses some risk that the product will be withdrawn from the market. Furthermore, patent owners can threaten customers and suppliers with patent lawsuits because patent infringement liability extends to every party who makes, uses, or sells a patented technology without permission, and sometimes to those who participate indirectly in the infringement.⁵ Furthermore, some of these costs persist after settlement.

Even simple delay can impose large business costs. Consider, for example, litigation against Cyrix, a startup firm that introduced Intel-compatible microprocessors.⁶ Intel, the dominant microprocessor maker, sued Cyrix and the suit lasted a year and a half. During that time, Cyrix had

⁴ See Mark Schankerman, *How Valuable is Patent Protection: Estimates by Technology Fields*, 29 RAND J. ECON. 77, 78, 95 (1998).

⁵ 35 U.S.C. § 271 (2006).

⁶ *Cyrix Corp. v. Intel Corp.*, 846 F. Supp. 522, 524 (E.D. Tex. 1994).

difficulty selling microprocessors to computer manufacturers, who were almost all customers of Intel as well, and who were reluctant to break ranks to go with a product that might be found to infringe. In the meantime, Intel responded by accelerating its development of chips that would compete against Cyrix's offerings. Ultimately, Cyrix won the lawsuit,⁷ but lost the war by losing much of its competitive advantage. Cyrix effectively lost the window of opportunity to establish itself in the marketplace; litigation exacted a heavy toll indeed.⁸

Although we explore the costs of litigation to both patent owners and alleged infringers in some detail, our chief interest is with the cost to alleged infringers. We choose this focus because innovators experience the patent system both as patent owners and as alleged infringers. Empirical methods that measure patent value by studying patent renewal or stock market valuation of patent portfolios account for the expected cost of enforcing patents through litigation.⁹ Unfortunately, there are no studies that quantify the negative impact of patent litigation cost on alleged infringers.

To the extent that costly patent litigation is primarily the result of inadvertent infringement—and we argue elsewhere that it is¹⁰—then the costs of defending against inadvertent infringement disincentivizes investment in innovation.¹¹ The risk of unavoidable infringement acts like a “tax” on innovation. We fear this tax has grown in recent years because we found that during the 1990s there was a dramatic increase in the hazard of patent litigation for publicly-traded firms.¹² More generally, one can view the costs of patent litigation as a negative “notice” externality imposed by the patent system.¹³

⁷ *Id.* at 541.

⁸ See generally James Bessen, et. al., *The Private and Social Costs of Patent Trolls*, REGULATION, Winter 2011-2012, at 26; Catherine Tucker, *Patent Trolls and Technology Diffusion* (Mass. Inst. of Tech., Working Paper 2011), available at <http://ssrn.com/abstract=1976593> (exploring how litigation by a patent troll affected the sales of medical imaging technology).

⁹ Nevertheless, it is useful to know how much patent value is eaten away by patent litigation, and what sort of reforms might reduce patent enforcement costs. Answers to those questions will have to wait for future research.

¹⁰ See James Bessen & Michael Meurer, *The Patent Litigation Explosion* 1, 9 (B.U. Sch. of Law, Working Paper Series, Law and Econ., Working Paper No. 05-18, 2005), available at <http://ssrn.com/abstract=831685> [hereinafter Bessen & Meurer, *The Patent Litigation Explosion*]; James Bessen & Michael Meurer, *Patent Litigation with Endogenous Disputes*, 96 AM. ECON. REV. 77, 81 (2006) [hereinafter Bessen & Meurer, *Patent Litigation with Endogenous Disputes*]; BESSEN & MEURER, *supra* note 2, at 192.

¹¹ These costs include the deadweight losses described above and also the settlement transfer from an innocent innovator/infringer to the patent owner.

¹² Bessen & Meurer, *The Patent Litigation Explosion*, *supra* note 10, at 17.

¹³ Peter S. Menell & Michael J. Meurer, *Notice Failure and Notice Externalities* 9-11 (B.U. Sch. of Law, Working Paper No. 11-58, 2012), available at <http://www.bu.edu/law/faculty/scholarship/workingpapers/documents/MenellP-MeurerM121611.pdf>.

The event study methodology has been used before to study litigation, beginning with Cutler and Summers in 1988¹⁴ in the context of merger litigation. Several papers have performed event studies of patent litigation—both of the initial filing event and of the terminating event, i.e., settlement, judgment, or verdict.¹⁵

These studies of initial filings, however, do not provide the best estimates from which to calculate the aggregate risk of infringement to the firms that perform R&D. They use small, selective samples and their estimates of wealth loss are not especially precise. Our contribution is to work with a much larger set of disputes. Our sample covers most patent lawsuits filed against U.S. public firms from 1984 through 1999—a sample responsible for the lion's share of R&D spending. Thus, our results are more precise and more representative of R&D-performing firms, permitting us to calculate a variety of cost and risk measures to inform policy. We find, in fact, that the estimates of wealth loss reported in some earlier studies appear to be overstated.

A key assumption of this literature is that the change in firm value that occurs around a lawsuit filing reflects investors' estimates of the direct and indirect effects of the lawsuit on the profits of the firm on average, and do not systematically reflect any unrelated information. We present evidence below that the revelation of unrelated information does not overstate our estimates for defendants in infringement suits and that, therefore, we may associate the loss in wealth with the effective total cost of litigation for defendants.

We find that alleged infringers lose about half a percentage point of their stock market value when sued for patent infringement. This corresponds to a mean cost of \$28.7 million in 1992 dollars (median of \$2.9 million), much larger than mean legal fees of about half a million dollars. In aggregate, infringement risk rose sharply during the late 1990s, exceeding \$16 billion in 1992 dollars for U.S. public firms. This amounts to 19% of these firms' R&D spending, a ratio that exceeds some estimates of the value of patents granted relative to R&D.

The next section describes the data and methods used for estimating cumulative abnormal returns. Section II reports average returns and some analysis of factors that affect returns. Section III calculates litigation cost,

¹⁴ David Cutler & Lawrence Summers, *The Costs of Conflict Resolution and Financial Distress: Evidence from the Texaco-Pennzoil Litigation*, 19 RAND J. ECON. 157, 159-64 (1988).

¹⁵ See generally Sanjai Bahagat et. al., *The Costs of Inefficient Bargaining and Financial Distress: Evidence from Corporate Lawsuits*, 35 J. FIN. ECON. 221, 245-46 (1994) [hereinafter Bahagat, *Costs of Inefficient Bargaining*]; Sanjai Bahagat et. al., *The Shareholder Wealth Implications of Corporate Lawsuits*, 27 FIN. MGMT. 5, 24-25 (1998) [hereinafter Bahagat, *Shareholder Wealth*]; Bruce Haslem, *Managerial Opportunism During Corporate Litigation*, 60 J. FIN. 2013, 2016-19 (2005); Josh Lerner, *Patenting in the Shadow of Competitors*, 38 J.L. & ECON. 463, 489-91 (1995); Glynn S. Lunney, Jr., *Patent Law, the Federal Circuit, and the Supreme Court: A Quiet Revolution*, 11 SUP. CT. ECON. REV. 1, 77-78 (2004).

Section IV calculates some broader measures of infringement risk, and Section V concludes.

I. DATA AND METHODS

A. *Data Sources*

Our research matched records from three data sources: lawsuit filings from Derwent's Litalert database, firm financial data from Compustat, and CRSP data on securities prices. In addition, we searched The Wall Street Journal's electronic archives to locate any articles announcing lawsuit filings or other events that might confound our analysis.

Using these sources, we constructed two main samples. The first, a small sample, only included lawsuits that identify public firms on both sides of the dispute. The second, a large sample, included all cases where the alleged infringer—defendant in an infringement suit or plaintiff in a declaratory action—but not necessarily the patentee litigant, was a publicly traded firm.

Our primary source of lawsuit filings information was Derwent's Litalert database, a database that has been used by several previous researchers.¹⁶ Federal courts are required to report all lawsuits filed that involve patents to the U.S. Patent and Trademark Office, and Derwent's data is based on these filings. Beginning with Derwent's data from 1984 through 2000, we removed duplicate records involving the same lawsuit, as identified by Derwent's cross-reference fields. We also removed lawsuits filed on the same day, with the same docket number, and involving the same primary patent. Sometimes, firms respond to lawsuits by filing counter-suits, perhaps involving other patents. Since our main focus is on initial disputes rather than lawsuit filings per se, we also removed filings made within 90 days of a given suit that involved the same parties.

The Derwent data does not distinguish between infringement and declaratory judgment suits. A firm threatened with an infringement suit can file a declaratory action seeking a judgment that the patent is invalid or not infringed. To classify each suit, we first identified whether the patent assignee of the main patent at issue matched a party to the suit. If the assignee matched a plaintiff, the suit was classified as an infringement suit. If the assignee matched a defendant, the suit was classified as a declaratory action. We matched the assignee for 83% of the suits and, of these suits,

¹⁶ Bessen & Meurer, *The Patent Litigation Explosion*, *supra* note 10, at 11; Jean O. Lanjouw & Mark Schankerman, *Protecting Intellectual Property Rights: Are Small Firms Handicapped?*, 47 J.L. & ECON. 45, 49 (2004); Rosemarie Ham Ziedonis, *Don't Fence Me In: Fragmented Markets for Technology and the Patent Acquisition Strategies of Firms*, 50 MGMT. SCI. 804, 815 (2004).

only 17% were declaratory actions.¹⁷ If the assignee did not match a party to the suit, then it was classified as an infringement suit because there are relatively few declaratory actions. This classification then allowed us to identify whether the subject firm was a “patentee litigant”—a plaintiff in an infringement suit or defendant in a declaratory action—or an “alleged infringer”—a defendant in an infringement suit, or plaintiff in a declaratory action.

To explore characteristics of firms involved in these lawsuits, we matched the listed plaintiffs and defendants to the Compustat database of U.S. firms from 1984 to 1999 that reported financials (excluding American Depository Receipts of foreign firms traded on U.S. exchanges). This data is based on merged historical data tapes from Compustat and involved an extensive process of tracking firms through different types of reorganization while eliminating duplicate records for firms—e.g., consolidated subsidiaries are listed separately from their parent companies.¹⁸

We matched the lawsuit data to the Compustat data by comparing the litigant names with all domestic firm names in Compustat as well as with a list of subsidiary names used in Bessen and Hunt.¹⁹ To check the validity and coverage of this match, we randomly selected a number of parties to suits and then checked them manually using various databases, including PACER, LexisNexis, the Directory of Corporate Affiliations, and LexisNexis M&A. Although we were unable to definitively identify all parties, the rate of false positives was not more than 3%—no more than 5 of 165 parties were found to have been falsely matched—and the rate of false negatives was no more than 7%. No more than 34 of 502 public companies were not matched. Finally, we matched the Compustat firms to the CRSP file of daily security prices.

We identified 2,648 suits with sufficient data on alleged infringers, some with multiple alleged infringers, for a total of 2,887 events in our large sample. We also selected all lawsuits where we could identify at least one party on each side as a publicly listed firm. This left us with a sample

¹⁷ These numbers are quite similar to findings by Moore in 2000 and Lanjou & Schankerman in 2001. See Jean O. Lanjou & Mark Schankerman, *Enforcing Intellectual Property Rights* 1-44 (Nat'l Bureau of Econ. Research, Working Paper No. 8656, 2001), available at <http://www.nber.org/papers/w8656>; Kimberly Moore, *Judges, Juries and Patent Cases—An Empirical Peek Inside the Black Box*, 99 MICH. L. REV. 365, 404 (2000).

¹⁸ This work was conducted by Bob Hunt and Annette Fratantaro at the Federal Reserve Bank of Philadelphia for an earlier project and we thank them for graciously sharing it with us. James Bessen & Robert M. Hunt, *An Empirical Look at Software Patents*, 16 J. ECON. & MGMT. STRATEGY 156, 158-59 (2007).

¹⁹ *Id.* A software program identified and scored likely name matches, taking into account spelling errors, abbreviations, and common alternatives for legal forms of organization. The matches were then manually reviewed and accepted or rejected. Note that this match is based on the actual parties to the litigation, not the original assignee of the patent at issue.

of 750 plaintiffs and 747 defendants in lawsuits where public firms were parties on both sides.

Table 1 shows our samples' summary statistics and further details from a closely related sample are reported in another Bessen and Meurer paper.²⁰ Parties to patent lawsuits tend to be larger than average firms with large R&D budgets. Moreover, our large sample captured the bulk of patent litigation against R&D performers. In 1999, U.S. public firms in Compustat spent \$150 billion on R&D, while total industrial R&D spending reported by the National Science Foundation was \$160 billion.²¹ Aside from under-reporting issues, our large sample constitutes a comprehensive sample with which we can obtain a lower bound measure of the aggregate risk of infringement to R&D performers.

Table 1. Summary Statistics

	Matched Sample				All Alleged Infringers	
	Patentee Litigants		Alleged Infringers		Mean	Median
	Mean	Median	Mean	Median		
Sales (\$ million)	7,020.4	1,267.7	6,186.7	1,022.7	8,604.0	1,368.1
Employees (1000s)	40.2	9.2	36.1	6.7	46.3	9.3
R&D/Sales	9.4%	5.4%	18.9%	5.3%	13.9%	5.0%
No R&D reported	6.1%		9.0%		18.4%	
No. observations	771		720		2887	

Finally, we checked each lawsuit in the small sample against The Wall Street Journal archive to identify suits that were announced in the Journal within one month of the filing date, and to identify possible confounding news about either party to the suit within one week of the filing date. In Section III, we discuss a supplemental dataset of lawsuits that reports legal fees.

²⁰ Bessen & Meurer, *Patent Litigation Explosion*, *supra* note 10, at 11-13.

²¹ There were important differences in the scope of what was included in these two measures. Nevertheless, they suggest that public firms account for the lion's share of R&D spending.

B. *Estimating Cumulative Abnormal Returns*

We used event study methodology²² to estimate the impact of filing a lawsuit on a firm's value. In particular, we used the dummy variable method described by Michael Salinger.²³ This assumes that stock returns follow a market model,

$$(1) \quad r_t = \alpha + \beta r_t^m + \epsilon_t,$$

where r_t is the return on a particular stock at time t , r_t^m is the compounded return on a market portfolio, and ϵ_t is a stochastic error. If an event like a lawsuit filing occurs on day T , then there may be an "abnormal return" to the particular stock on that day. This can be captured using a dummy variable,

$$(2) \quad r_t = \alpha + \beta r_t^m + \delta I_t + \epsilon_t,$$

where I_t equals 1 if $t=T$ and 0 otherwise. Equation (2) can be estimated using Ordinary Least Squares (OLS) for a single event. In practice, this equation is estimated over the event period, as well as over a sufficiently long pre-event window. In this paper, we used a 200 trading-day pre-event window.²⁴ The coefficient estimate of δ obtained by this procedure was then an estimate of the abnormal return on this particular stock. For different stocks, the precision of the estimates of δ will vary depending on how well equation (2) fits the data. The estimated coefficient variance from the regression provided a measure of the precision of the estimate of the abnormal return.

We wanted to obtain a representative estimate of the abnormal returns from lawsuit filings for multiple stocks under the assumption that these represent independent events and that they share the same underlying "true" mean. Previous papers estimating abnormal returns from patent lawsuits have simply reported unweighted means for the group of firms. Although the unweighted mean is an unbiased estimator, it is not efficient. Since we are concerned with obtaining the best estimate to use in policy calculations, and not just testing the sign of the mean, we used a weighted mean to esti-

²² Craig A. Mackinlay, *Event Studies in Economics and Finance*, 35 J. ECON. LIT. 13, 14-16 (1997).

²³ Michael Salinger, *Standard Errors in Event Studies*, 21 J. FIN. & QUANTITATIVE ANALYSIS 39, 41-42 (1992) (showing that this model is mathematically equivalent to the widely-used OLS market model described in Brown and Warner); see also Stephen J. Brown & Jerold B. Warner, *Using Daily Stock Returns: The Case of Event Studies*, 14 J. FIN. ECON. 3, 16-17 (1985).

²⁴ We also ran regressions with a 180 day pre-event window that ended 30 days before the lawsuit filing. Cumulative abnormal returns were very close to those with a 200 day window that lasted up to the day before the event window.

mate the “average abnormal return,” where the weight for each observation is proportional to the inverse of the variance of the estimate of δ for that firm.²⁵

When we test our means against the null hypothesis that the true mean is zero, we report both the significance of t -tests using the weighted mean as well as the significance of the Z statistic,²⁶ a widely used parametric test of significance that incorporates the variation in precision across events.²⁷ In any case, the significant test results are relatively similar, as are those of some nonparametric tests.

As Salinger²⁸ notes, this procedure assumes that the returns for each event are independent of each other. However, when there are multiple defendants in a suit, returns may be systematically related. For example, one defendant may be a supplier to another, or two defendants may be unequal rivals. Thus, for the 188 lawsuits in the large sample with multiple defendants, we estimated the returns for the defendants to each suit jointly.

Finally, equation (2) describes the abnormal return for a single day. It is straightforward to design dummy variables to estimate a “cumulative abnormal return” (CAR) over an event window consisting of multiple consecutive days. In the following, for instance, if the suit is filed on date $t=T$, then we may use a window from day $T-1$ to $T+24$.

C. *The Event*

This paper also differs from previous research in the nature of the events we study. Previous studies have used the announcement of the lawsuit in a newspaper or wire service as the event. Instead, we use the filing of the lawsuit. This may seem to be a minor difference, but it is significant for two reasons.

First, at the time of our sample, most patent lawsuits were not announced in newspapers or wire service reports at all. Various factors may influence whether a lawsuit is announced or not. Firms may choose to issue a press release or not. The Securities and Exchange Commission (SEC) requires reporting of major lawsuits in quarterly and annual filings but lawsuits will be reported separately only if they materially affect the profits of

²⁵ In any case, we find that for our entire sample, the weighted mean is quite close to the unweighted mean and also to the median. There are significant differences, however, in the averages for subsamples.

²⁶ See Peter Dodd & Jerold B. Warner, *On Corporate Governance: A Study of Proxy Contests*, 11 J. FIN. ECON. 401, 417-22, 425-28, 430-34, 436 (1983).

²⁷ See Lisa A. Kramer, *Alternative Methods for Robust Analysis in Event Study Applications*, ADVANCES IN INV. ANALYSIS & PORTFOLIO MGMT., 2001, at 1, 10 (using the Z statistic is a joint test of the individual firm t -tests).

²⁸ Salinger, *supra* note 23, at 39-42.

the firm. Accordingly, news sources may not report all lawsuits even if the firms issue press releases.

We took a random sample of patent lawsuits against U.S. public firms and searched LexisNexis for news stories that mention the lawsuits within one month of the filing date, both before and after. We found that only 19% of the lawsuits were mentioned in the Dow Jones Newswire, one of the most comprehensive reporting services; only 7% were mentioned in *The Wall Street Journal*, the source used in several of the previous studies. Since one of our objectives is to tally the combined risk of lawsuits for public firms, clearly we cannot obtain comprehensive estimates by relying solely on announced lawsuits.

Moreover, announced lawsuits are a select group that may be qualitatively different from other lawsuits. That is, samples of announced lawsuits may suffer from sample selection bias. In order to test this, we performed a series of Probit regressions in our small sample on whether a lawsuit was reported in *The Wall Street Journal*.²⁹ Among other things, we find that the probability of a *Wall Street Journal* announcement is strongly correlated with the defendant firm's stock market beta. This may reflect the editorial judgment of *The Wall Street Journal* that certain lawsuits are more newsworthy and more likely to affect a defendant's stock price. Alternatively, perhaps word of the lawsuit is already affecting the defendant's stock price. This, in turn, suggests that estimates made on a sample of announced lawsuits may have abnormal returns with a larger absolute magnitude than those from a more representative sample.

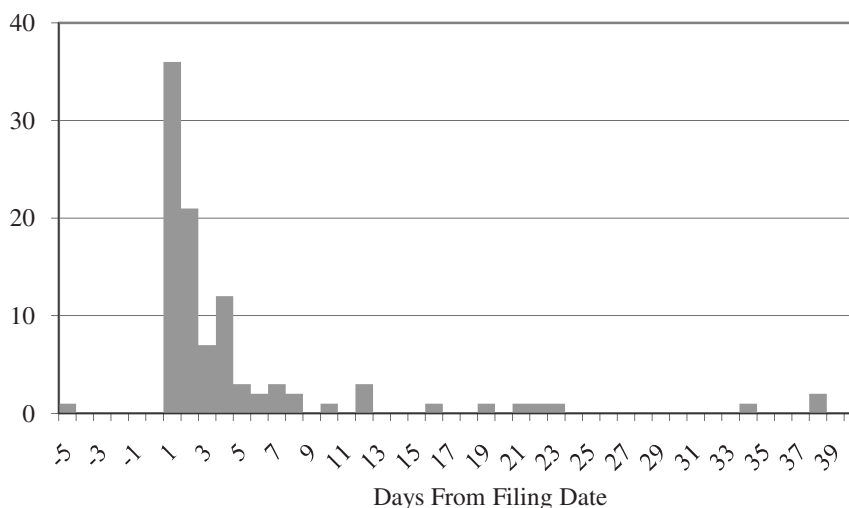
Below, we compare estimates of abnormal returns from samples of lawsuits announced in *The Wall Street Journal* with estimates from our comprehensive sample. We find that our estimates from the announced sample are quite similar to those reported in the previous literature. However, estimates from the previous literature are substantially larger in absolute magnitude than those for our comprehensive sample, suggesting considerable sample selection bias.

On the other hand, our estimates may be understated for another reason: investors may not receive news of the lawsuit within an event window around a filing date. With an announcement in a newspaper or major newswire, we can be reasonably sure that investors hear the news of the lawsuit within a day or two of the announcement. But we cannot be sure that investors hear the news about a legal filing in a district courthouse. Indeed, depending on how long it takes to serve papers, the defendant may not be aware of the lawsuit for a day or so after the filing date. In other words, news of an unannounced patent lawsuit filing may leak out more slowly, and investors may not learn of a lawsuit within a specified event window. This is particularly true for many of the small firms in our sample.

²⁹ See *infra* Appendix.

Because the mean CAR is often smaller than the bid-ask price gap for lightly traded stocks, it might be particularly difficult to make money arbitraging these securities. The money made in arbitrage depends on the volume of the trade, and attempts to arbitrage stocks with little “float” tend to make the price move against the arbitrageur. If profits are small, it will not pay to invest in information gathering activities needed to obtain news of patent lawsuits.

Figure 1. Frequency of The Wall Street Journal Stories Relative to Court Filing Date



We see evidence of this slower diffusion of information in the lawsuits that were announced in The Wall Street Journal. Figure 1 displays the frequency of these news stories relative to the actual court filing date. Event studies based on public announcements typically use an event window of two or three days, often occurring one day before the announcement. Although many lawsuits are announced within two days of filing, such a small event window around a *filing date* would clearly miss a very large share of lawsuit announcements. Moreover, it seems likely, given the role of stock market beta in the likelihood of a Wall Street Journal article, that the lawsuits that are announced within a few days of the filing may be qualitatively different from those of which the news leaks out more slowly and are either announced later or not announced at all. Indeed, we find evidence within our data that stocks with beta above 1 react to the filing faster than lower beta stocks.³⁰ In order to have representative and comprehensive estimates,

³⁰ At day 2, the higher beta stocks for defendant firms have a CAR that is significantly lower than the CAR for lower beta stocks (at the 5% level) and the lower beta CAR is not significantly different from zero. At day 24, the CARs for these two groups are not significantly different, but both are signifi-

we use a longer event window than would be appropriate in an announcement event study. Specifically, we use a 25 day window (from $T-1$ to $T+24$), which, based on the data in Figure 1, should capture 96% of the announced events and, we hope, a large share of the unannounced filings. We show some CARs from shorter windows in the Appendix.

There are two possible concerns with using a longer window. First, the longer window introduces more “noise” into the estimation, reducing precision and possibly attenuating the estimates. This is not such a significant concern, however, because we have a much larger sample size than earlier studies, and our estimates are reasonably precise, although they may be slightly attenuated. Second, research on long-horizon event studies—studies with *multi-year* event windows—find that certain biases arise for a variety of reasons.³¹ However, it seems highly unlikely that these concerns can exert a substantially greater influence in a 25 day window than they exert in a three day window.

In summary, restricting the study to events announced in news services likely introduces substantial sample selection bias. Our estimates, based on a larger window following the filing of the lawsuit, are smaller, although they might be biased toward zero.

II. EMPIRICAL FINDINGS

A. *Estimates of Cumulative Abnormal Returns*

Since previous studies have used samples where parties on both sides of a lawsuit were public firms and the suits were reported to newspapers or wire services, we begin by exploring a sub-sample. Table 2 shows estimates of CARs for just those suits from our small matched sample that were reported in The Wall Street Journal. In Table 2, we exclude suits that had a potentially confounding news story in The Wall Street Journal within a month of the filing date. Two previous studies have reported on event study estimates on announcements of patent lawsuits filings. Bhagat et al.

cantly different from zero. One explanation for the faster speed of diffusion for high beta stocks is that the opportunities for investors to make returns from the information about the lawsuit filing are relatively greater for these stocks.

³¹ Brad M. Barber & John D. Lyon, *Detecting Long-Run Abnormal Stock Returns: The Empirical Power and Specification of Test Statistics*, 43 J. FIN. ECON. 341, 341-72 (1997); S. P. Kothari & Jerold B. Warner, *Measuring Long-Horizon Security Price Performance*, 43 J. FIN. ECON. 301, 301-39 (1997). These reasons include: (1) with a long window, the composition of the market index may change with the addition of new entrants or from rebalancing; (2) compounding of returns leads to a highly skewed distribution; (3) not all firms survive to the end of a long event window; and (4) the market model or its variance may change or may be sensitive to specification errors over long windows. We find that our measured returns are not highly skewed, and there are few cases of firms failing to survive the event window.

examined lawsuits filed between 1981 and 1983 (51 plaintiffs and 33 defendants) and Lerner obtained estimates for 26 biotech lawsuits from 1980 to 1992.³² To maintain consistency with the previous literature, in Table 2 (but not in Table 3), we report simple unweighted means of CARs.³³ The mean and median values are reported for two different event windows: one around The Wall Street Journal publication date and the other, a longer window around the actual suit filing date reported in court records. Notably, these dates occasionally differed significantly.

Table 2. Cumulative Abnormal Returns from Suits Announced in The Wall Street Journal, 1984–1999

Event window	WSJ article T-2 to T+1	Suit filing T-1 to T+24	Bhagat et al. (1998)
<u>Patentee Litigant (Plaintiff)</u>			
mean	-0.3%	-0.1%	-0.31%
median	0.0%	0.9%	
no. of observations	86	86	
<u>Alleged Infringer (Defendant)</u>			
mean	-2.6%	-1.8%	-1.50%
median	-1.4%	-1.9%	
no. of observations	82	82	
<u>Combined (matched parties)</u>			
mean	-2.6%	-2.5%	
median	-1.8%	-0.5%	
no. of observations	80	80	
Addendum: mean combined abnormal returns			
Bhagat et al. (1994)		-3.13%	
Lerner (1995)		-2.0%	

Note: Events with possibly confounding news are excluded. Average cumulative abnormal returns are simple unweighted means.

³² Bhagat et al., *Shareholder Wealth*, *supra* note 15, at 20; Lerner, *supra* note 15, at 471. Bhagat et al. (1998) included the data from the Bhagat et al. 1994 paper, so we did not list that data separately. Lerner searched The Wall Street Journal as well as news wire services for announcements. The other studies limited their use only to articles in The Wall Street Journal.

³³ For this reason, this table does not report standard errors or significance tests.

Consistent with most of the previous literature on litigation, we found that patentee litigants do not show a positive response to a lawsuit filing. Bhagat et al. (1998) reported a CAR of -0.31%, and we found a similar value.³⁴ For defendants—alleged infringers—we found a substantial loss in market value of around 2%. Bhagat et al. reported a loss of 1.5%.³⁵ For the combined loss of wealth, we found a mean of 2.5 – 2.6%, although with smaller median values. Bhagat et al. (1994) reported a mean loss of 3.13% and Lerner (1995) reported a mean loss of 2.0%.³⁶ All three results are broadly similar and quite substantial. Lerner reported a mean absolute loss of shareholder wealth of \$67.9 million, a median loss of \$20 million.³⁷ In general, there does not appear to be a major difference between the results reported in the event window around The Wall Street Journal publication date and the longer window around the filing date.

As noted above, estimates for this sub-sample may be unrepresentative of most patent litigation, however, because most lawsuits are not reported in The Wall Street Journal. Table 3 reports cumulative abnormal returns for all lawsuits in the matched sample, found at the top of the table, as well as those for the large sample, which are found at the bottom of the table. The base result for the matched sample used a 25 day event window ($T-1$ to $T+24$) and excluded lawsuits when we identified possibly confounding events. The table also reports CARs for suits that were positively identified as infringement suits—the plaintiff was the patent assignee—and for a sample that included lawsuits with possibly confounding news events. The reported means and standard errors use weights based on the variance of the dummy variable coefficient in the event regression. Several results stand out.

First, the estimated percentage losses for alleged infringers are substantially less than those for lawsuits reported in The Wall Street Journal in Table 2. We cannot tell, however, whether the percentage loss estimates in the Journal are larger because of a selection effect or because of the greater information conveyed by publication in the Journal. Even though some learning takes place, we suspect that in most lawsuits, investors remain relatively uninformed compared to those cases where an announcement is published in The Wall Street Journal. The SEC requires reporting of major lawsuits in quarterly and annual filings, but lawsuits will be reported separately only if they materially affect the profits of the firm. For a handful of suits, we checked published sources and typically found no mention of the suit. For this reason, estimates for the non-Journal sample should be inter-

³⁴ Bhagat et al., *Shareholder Wealth*, *supra* note 15, at 18.

³⁵ *Id.*

³⁶ Bhagat et al., *Costs of Inefficient Bargaining*, *supra* note 15, at 230; Lerner, *supra* note 15, at 471.

³⁷ Lerner, *supra* note 15, at 471.

preted as lower bound estimates of defendant firms' loss of wealth—significant numbers of investors likely became informed about the suit either after our event window or, if there were pre-filing interactions, before.

Second, patentee litigants/plaintiffs appear to suffer some losses as well. These losses are smaller than those for alleged infringers/defendants, but they are statistically significant.³⁸ This is consistent with previous research and it indicates that lawsuits do not represent simple transfers of wealth on average. Instead, there is dissipation of wealth to consumers, to rivals or to deadweight loss.

Finally, the magnitudes of returns for definite infringement suits are generally larger than for those of all suits, and they show a higher level of statistical significance. This may be because among those cases where we could not match the patent to one of the parties, some plaintiffs are mistakenly classified as defendants and vice versa. Or it could be due to the fact that declaratory actions may be more likely when the stakes at issue are smaller or that alleged infringers have an advantage at choosing a friendly court when they file a declaratory action.

The bottom of Table 3 reports results for our large sample. The CARs for alleged infringers are similar to those obtained from the smaller sample—a loss of 0.5% to 0.6%—but here they are statistically significant at the 1% level, except for those lawsuits involving multiple defendants.

³⁸ It might seem puzzling that the average market response when a patent holder files a lawsuit is negative. Individual rationality implies that the patent holder only files lawsuits that have positive expected value. If this is the only relevant information, then plaintiff CARs should be positive. As we explain in more detail in Section III.B, the event of filing may reveal information to investors about more than just the lawsuit. Filing might reveal private information that the patent holder's patent is stronger than investors believed, or that the patent holder has better technology or better entry prospects than investors believed. These possibilities provide additional reasons for why the patent holder's share value should rise with the filing of a lawsuit. In contrast, filing might reveal private information of patent weakness, or that a tacit industry agreement not to file patent lawsuits has broken down. These possibilities suggest share value should fall upon lawsuit filing. Thus, a negative CAR might be explained as follows: When a pharmaceutical firm files a patent suit, investors perceive the suit has positive expected value, but they also perceive that a key patent was not as strong as they thought and did not deter entry by a potential competitor. Alternatively, when a semiconductor firm files a patent suit, investors perceive that the suit has a positive expected value. Investors however, also perceive that the patent holder plans to exit the industry or has become less forward-looking for some reason, and the firm is therefore willing to deviate from a no-lawsuit equilibrium. Further research is required to resolve this puzzle.

Table 3. Cumulative Abnormal Returns

	Mean CAR	Median CAR	Robust Z Statistic	Observations
Sample: Matched Parties				
<u>Patentee Litigants</u>				
Base	-0.38% (0.30%)	0.00%	-1.51	667
Definite infringement suits	-0.63% (0.37%)*	-0.45%	-2.18*	412
<u>Alleged Infringers</u>				
Base	-0.62% (0.33%)*	-0.97%	-1.55	661
Definite infringement suits	-0.77% (0.42%)*	-0.83%	-1.70*	407
With possibly confounding events	-0.45% (0.31%)	-0.57%	-1.32	743
Sample: All alleged infringers				
Base	-0.50% (0.16%)**	-0.51%	-3.24**	2,887
Single defendants	-0.61% (0.18%)**	-0.54%	-2.94**	2,460
Multiple defendants	-0.01% (0.39%)	-0.39%	-1.38	427
Single defendants, definite infringement cases	-0.63% (0.27%)**	-0.42%	-2.37**	1,108

Note: Standard errors are in parentheses. A single asterisk indicates statistical significance at the 5% level; a double asterisk indicates significance at the 1% level. Average cumulative abnormal returns (CARs) are weighted means with weights proportional to the inverse of the estimated variance of each return. In the matched sample events, possibly confounding news are excluded, except where noted. The event window is twenty-five days ($T-1$ to $T+24$). Cumulative abnormal returns are estimated using OLS except for cases with multiple defendants (in the large sample), which are estimated jointly. The robust Z statistic is a joint test of the individual firm t statistics. Kramer, *supra* note 27.

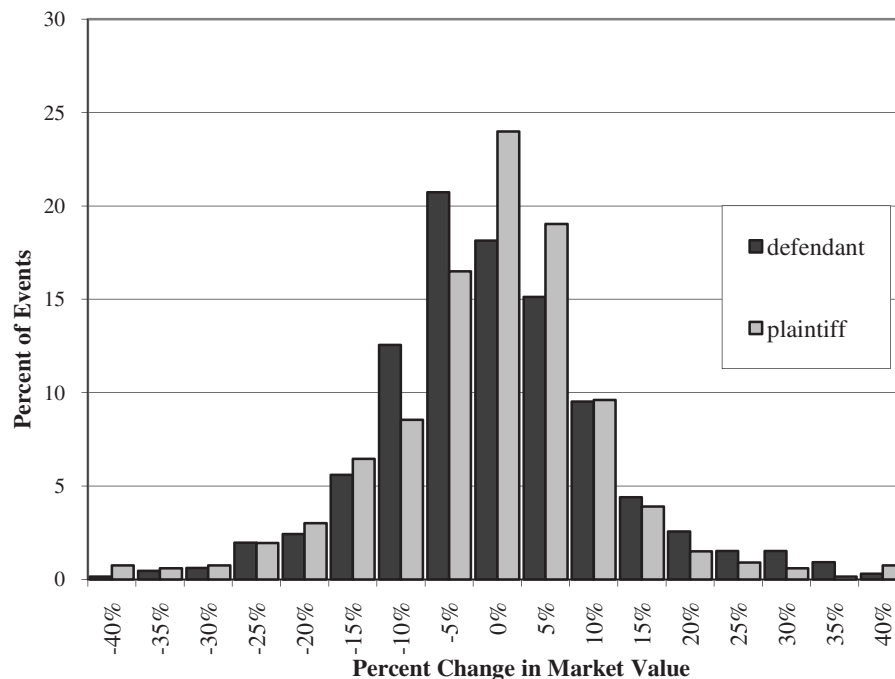
When multiple defendants are involved, the returns are negligible, suggesting that something is fundamentally different about these estimates. There are several possible explanations for this. It may be that suits naming multiple defendants are more frivolous, so that investors do not expect serious losses. Alternatively, some defendants may have been contractually indemnified, diluting the estimates. A higher percentage of defendants in lawsuits with multiple defendants are from retail and wholesale industries, suggesting that these suits more frequently involve downstream resellers

who have less at stake. Furthermore, costs may be shared among multiple defendants, reducing the individual firm costs.

The estimates in the lower portion of the table do not control for possibly confounding events. However, we find that excluding observations with possibly confounding events does not seem to substantially alter the mean estimated CARs in the top portion of Table 3—the matched parties sample. To check this further, we repeated the estimates for the large sample of all alleged infringers, but we terminated the pre-event window 30 days prior to the filing of the lawsuit. This made little difference in our estimates, suggesting that confounding events may add noise, but do not bias our estimates.³⁹

Figure 2 shows histograms for the cumulative abnormal returns for all lawsuits from the matched sample. The curve for alleged infringers/defendants clearly falls to the left of the curve for patentee litigants/plaintiffs, but both curves are quite diffuse. The distributions are significantly leptokurtic—with a kurtosis of 7.2 and 9.7 for plaintiffs and defendants, respectively—meaning that they have long tails. This suggests that outliers may be influential. To make sure that our results are not driven by outliers, we also conducted non-parametric tests—the binomial probability test and the Wilcoxon signed rank test—on the large sample and several sub-samples. All of these tests rejected the null hypothesis of a CAR being zero at either the 5% or the 1% level of statistical significance. In addition, the close correspondence between the means and the medians suggests that our mean estimates for alleged infringers are representative.

³⁹ For example, the estimate for single defendants was 0.608% (0.176%) for the full 200 day pre-event window and 0.609% (0.178%) for the truncated window.

Figure 2. Histograms of Cumulative Abnormal Returns

Finally, the CARs we measure may reflect some sort of temporary over-reaction by investors. For instance, suppose that nervous investors sell heavily immediately upon news of the lawsuit, driving the stock price below what is warranted given the firm's expected profits. Later, more savvy investors, recognizing the low price, buy up shares, restoring the price to a level that more accurately reflects potential profits. In this case, our initial measure of the CAR will be too negative. But if this were the case, then we would expect that a longer observation window would make the CARs less negative as savvy investors entered after the initial over-reaction. However, the evidence shows that CARs become *more* negative with a longer window. This suggests, instead, that the CARs we measure do not reflect a temporary over-reaction.

Another possible bias might arise if investors felt that the lawsuit greatly affects the variability of the expected profits—that is, if the lawsuit increases the uncertainty of future profits. In this case, investors might demand a greater risk premium while the lawsuit is underway. When the suit is resolved, by settlement or adjudication, the original risk premium should return. In this case as well, the CARs we measure might not accurately reflect the long-term prospects of the firm—a portion of the drop in stock price might be due to the temporarily greater risk premium instead. How-

ever, some studies have looked at what happens when patent lawsuits are resolved.⁴⁰ These studies do not find positive CARs when a lawsuit is settled; in fact, one study found negative CARs.⁴¹ Thus, this finding implies that lawsuits do not significantly alter investors' risk premiums for the defendant firms. We conclude that the evidence we found does not indicate that our CARs reflect temporary changes in investor sentiment or risk premiums; instead, they likely reflect permanent changes in investors' valuations of the firm.

B. *Factors affecting Abnormal Returns*

Tables 4 and 5 explore factors that might influence the magnitude of investors' reactions to lawsuit filings by comparing means of different sub-groups. We tested differences in the means of different sub-groups using one-tailed *t*-tests, allowing unequal variances between the sub-groups and calculating the degrees of freedom using Satterthwaite's approximation.⁴² We conducted these comparisons both for the subject firm's characteristics as well as for characteristics of its opposing party in the lawsuit. We also ran regressions with various combinations of the variables in Table 4, or continuous equivalents, on the right hand side. However, given the noisiness of our data, little conclusive evidence could be drawn from these regressions and where significant results were found, they matched the results found with simple *t*-test comparisons of means.

Table 4. Differences in Mean CARs by Characteristics

Firm characteristic	Sample: Matched Parties	
	Alleged Infringer	Patentee Litigant
Employees < 500	-3.20% (2.32%)	-3.18% (2.45%)
R&D/Sales > .15	0.22% (2.16%)	-0.53% (1.22%)
Total liabilities/Total Assets > .5	1.40% (0.87%)	-2.35% (0.75%)**
Capital/Employee > \$100,000	-0.02% (0.93%)	-1.02% (0.74%)
Current Assets/current liabilities < 1.5	0.94% (1.00%)	-1.91% (0.87%)*
Newly public firm	-0.94% (1.78%)	-1.92% (2.56%)

⁴⁰ Bhagat et al., *Shareholder Wealth*, *supra* note 15, at 5, 10; Haslem, *supra* note 15, at 2019.

⁴¹ Bhagat et al., *Shareholder Wealth*, *supra* note 15, at 16.

⁴² F.E. Satterthwaite, *An Approximate Distribution of Estimates of Variance Components*, 2 BIOMETRICS BULL. 110, 110-14 (1946).

Rival characteristic

Employees < 500	1.06% (1.19%)	-1.37% (1.07%)
R&D/Sales > .15	0.23% (1.62%)	0.81% (0.97%)
Total liabilities/Total Assets > .5	-0.15% (0.86%)	-0.35% (0.80%)
Capital/Employee > \$100,000	-0.99% (0.95%)	1.02% (0.74%)
Current Assets/current liabilities < 1.5	1.69% (1.11%)	1.19% (0.86%)
Newly public firm	3.77% (1.51%)**	0.32% (1.05%)

Other Characteristics

Year > 1989	-0.15% (0.82%)	0.09% (0.77)%
Firms in same SIC4 primary industry	2.67% (1.16%)**	-0.11% (0.78%)

Note: Standard errors are in parentheses. A single asterisk indicates that the difference is statistically significant at the 5% level; a double asterisk indicates significance at the 1% level (one-tailed test allowing unequal variances and using Satterthwaite's calculation for degrees of freedom). Average cumulative abnormal returns are weighted means, where weights are proportional to the inverse of the estimated variance of each return. Comparisons are for cases where infringement is known and no possibly confounding events have been found.

For patentee litigants, we found that firms with high liabilities relative to assets, and to a lesser extent, firms with high current liabilities to current assets, have much more negative returns from initiating lawsuits. One explanation is provided by Haslem, who observed that on average, lawsuit settlements, including patent settlements, are associated with a decline in firm value.⁴³ Following Jensen and Meckling, Haslem argued that, from the perspective of shareholders, poorly governed firms will tend to settle lawsuits too soon because early settlement allows managers to expend less effort.⁴⁴ Firms with low debt have more leeway for managerial discretion.⁴⁵ Haslem found that these firms experience greater declines in value from settlement.⁴⁶ By similar logic, firms with low debt may have more discretion about which lawsuits to file. Therefore, they may choose to file only the most profitable lawsuits while managers in more debt-laden companies

⁴³ Haslem, *supra* note 15, at 2025, 2027.

⁴⁴ *Id.* at 2014, 2040; Michael C. Jensen & William H. Meckling, *Theory of the Firm: Managerial Behavior, Agency Costs, and Ownership Structure*, 3 J. FIN. ECON. 305 (1976).

⁴⁵ Haslem, *supra* note 15, at 2039-40.

⁴⁶ *Id.*

may be driven to file more marginal lawsuits, leading to relatively lower CARs.

Another explanation might arise if some industries have a “mutual forbearance” repeated game type equilibrium—where firms mutually avoid suing each other because they fear retaliatory suits. However, a failing firm may have limited future prospects, hence little to fear from future retaliation. Thus, failing firms with high liabilities may be more likely to initiate suits, including less profitable suits.

For alleged infringers, we found five statistically significant differences. First, if the parties to the lawsuit are in different industries, then the alleged infringer suffers a substantially larger loss, which is statistically significant at the 1% level. Suits from outside the industry may be more of a surprise to investors and may be more indicative of inadvertent infringement. Alternatively, when disputes occur within a narrow industry, the parties may have greater latitude to craft a settlement that benefits both jointly, including even collusive settlements.

Table 5. Differences in Mean CARs by Firm Characteristics

Sample: All alleged infringers	
Employees < 500	-1.70% (0.92%)*
R&D/Sales > .15	-1.79% (0.80%)*
Total liabilities/Total Assets > .5	0.05% (0.33%)
Capital/Employee > \$100,000	-0.26% (0.44%)
Current Assets/current liabilities < 1.5	0.11% (0.34%)
Year > 1989	-0.56% (0.32%)*
Patentee is public firm	-0.12% (0.35%)
Industry	
SIC = 28 (chemicals, inc. pharma)	-0.41% (0.41%)
SIC = 35, 36, 73 (electronics, computer, sw)	0.06% (0.38%)
Other manufacturing	0.16% (0.33%)

Note: Standard errors are in parentheses. A Single asterisk indicates difference is statistically significant at the 5% level; a double asterisk indicates significance at the 1% level (one-tailed test allowing unequal variances and using Satterthwaite’s calculation for degrees of freedom). Average cumulative abnormal returns are weighted means, where weights proportional to the inverse of the estimated variance of each return.

Second, if the patentee litigant is a newly public firm, the alleged infringer makes out better. This might be because newly public firms are less able to pursue sustained litigation, posing less of a threat to the alleged in-

fringer. Or, perhaps, a suit by an entrant firm provides a signal that the technology may be more profitable than investors previously realized.⁴⁷

The remaining three differences from the large sample, shown in Table 5, are statistically significant at the 5% level. First, small firms seem to have substantially more negative returns. This result appears robust to alternative cutoff points below 500 employees, but we found no significant variation in returns among firms larger than 500. One explanation for this is that legal costs are relatively higher for small firms, creating a “floor” on the costs of litigation. Second, we found limited evidence that R&D intensive firms suffer more negative returns. However, this result seems sensitive to the specific cutoff used. Finally, we also found some evidence of worse returns during the 1990s compared to the 1980s. Notably, the lower returns for alleged infringers do not appear to be matched by greater returns to patentee litigants (top of Table 5). In other words, the evidence of greater losses does not suggest a greater transfer of wealth to patent holders.

III. THE COSTS OF PATENT LITIGATION

A. *Legal Costs*

We first looked at attorneys’ fees in patent litigation using supplemental data we collected from legal records.⁴⁸ We then estimated the total costs of litigation to alleged infringers based on our event study estimates.

Public documents in certain U.S. patent lawsuits record attorneys’ fees because American patent law gives judges the discretion to shift fees in exceptional cases. Patentees usually get fee awards based on a finding of willful infringement, and alleged infringers usually get fee awards based on a finding that the patent suit was frivolous or vexatious. In searching Westlaw for all patent cases from 1985 to 2004 that discussed fee-shifting, we found 352 cases in which one of the parties requested fees (about 100 patent cases go to trial per year). The request was granted in 137—38.9%—of these cases. From this set of 137 cases, we were able to determine the magnitude of the fees in 87 cases—63.5% of awards—from either judicial opinions or from documents filed by the parties available through the PACER system.

Table 6 shows the median and mean amounts of the fee awards in millions of dollars with 1992 dollars as the base. Mean fees for cases that went through trial were \$1.04 million for patentee litigants and \$2.46 million for alleged infringers. For cases that were decided prior to trial, the mean fees

⁴⁷ See Section III.B for a discussion about how a suit by an entrant firm provides a signal that the technology may be more profitable than investors previously realized.

were \$950,000 for patentee litigants and \$570,000 for alleged infringers.⁴⁹ Median values tend to be smaller because the distribution is skewed. In the most extreme case, a \$26 million fee was awarded to Bristol-Myers Squibb in conjunction with a successful defense against a pharmaceutical patent suit brought by Rhone-Poulenc.⁵⁰ The next largest award was about \$7 million.

**Table 6. Attorneys' Fees Awarded in Patent Lawsuits
(in millions of year 1992 dollars)**

	Mean	Median	Observations
<u>Patentee Litigant</u>			
Summary Judgment	.95	.40	8
Verdict	1.04	.78	51
<u>Alleged Infringer</u>			
Summary Judgment	.57	.30	10
Verdict	2.46	.98	18

Our fee-shifting data is in line with survey information collected by the American Intellectual Property Law Association (AIPLA).⁵¹ AIPLA asked patent litigators to estimate the fees associated with patent lawsuits under six different scenarios.⁵² Specifically, the survey question divided cases into three different intervals based on stakes.⁵³ The survey then asked for estimates for cases that concluded at the end of discovery, and for cases that reached trial. Their 2001 report indicated that the estimated cost through trial was \$499,000 when the stakes were less than \$1 million, \$1.499 million when the stakes were between \$1 million and \$25 million,

⁴⁹ We included cases that ended in summary judgments, one case that settled, one case that was a default judgment, and one case that ended in a motion to dismiss.

⁵⁰ *Bristol-Myers Squibb Co. v. Rhone-Poulenc Rorer, Inc.*, 2002 U.S. Dist. LEXIS 13706, at *49 (S.D.N.Y. July 25, 2002), *aff'd* 326 F.3d 1226, 1233 (Fed. Cir. 2003).

⁵¹ AM. INTELL PROP. LAW ASS'N, AIPLA REPORT OF THE ECONOMIC SURVEY 2001 (2001).

⁵² *Id.* at 14.

⁵³ *Id.* at 16.

and \$2.992 million when the stakes were over \$25 million.⁵⁴ The estimated cost through discovery was \$250,000 when the stakes were less than \$1 million, \$797,000 when the stakes were between \$1 million and \$25 million, and \$1.508 million when the stakes were over \$25 million.⁵⁵

The expected legal cost associated with filing a patent lawsuit depends on the frequency of each of the different ways a lawsuit may be terminated. Kesan and Ball analyze patent lawsuit termination data available from the Administrative Office of the Federal Judiciary.⁵⁶ After examining 5,207 lawsuits filed in 1995, 1997, and 2000, they found that most cases terminate short of trial, summary judgment, or through other substantive court rulings.⁵⁷

In particular, 4.6% of lawsuits reached trial, 8.5% of lawsuits terminated with a summary judgment, dismissal with prejudice, or confirmation of an arbitration decision, and the remaining 86.9% of cases terminated earlier in the process.⁵⁸

Kesan and Ball constructed the following two proxies for legal fees in patent lawsuits: number of days until the suit terminates, and number of documents filed.⁵⁹ Their data showed that suits that go to trial last about 1.5 times as long as suits that end with a summary judgment, and suits that end with a summary judgment last about 1.5 times as long as all other suits.⁶⁰ Further, their data showed that suits that go to trial generate about 2.5 times as many documents as suits that end with a summary judgment, and suits that end with a summary judgment generate about 2.5 times as many documents as all other suits.⁶¹ Assuming that the expected legal cost in a suit that ends before summary judgment is one-half of the cost of a suit that reaches summary judgment, then the estimated amount for the alleged infringer is \$409,000 and \$541,000 for the patentee, as shown in Table 6.⁶² A similar calculation using AIPLA data for stakes between \$1 million and \$25 million yields an estimate of \$483,000.

⁵⁴ *Id.* at 84. These estimated cost through trial increased substantially in the 2003 and 2005 AIPLA reports. AM. INTELL PROP. LAW ASS'N, AIPLA REPORT OF THE ECONOMIC SURVEY 2005 (2005); AM. INTELL PROP. LAW ASS'N, AIPLA REPORT OF THE ECONOMIC SURVEY 2003 (2003).

⁵⁵ AM. INTELL PROP. LAW ASS'N, *supra* note 51, at 84-85. The AIPLA estimate of costs through discovery should be larger than the fees shifted at the summary judgment stage to the extent that discovery continues after summary judgment. *Id.*

⁵⁶ Jay P. Kesan & Gwendolyn G. Ball, *How are Patent Cases Resolved? An Empirical Examination of the Adjudication and Settlement of Patent Disputes*, 84 WASH. U. L. REV. 237, 238 (2006).

⁵⁷ *Id.* at 310-12.

⁵⁸ *Id.*

⁵⁹ *Id.*

⁶⁰ *Id.*

⁶¹ *Id.* at 280, 282, 285 (we derive these ratios from Tables 10-12).

⁶² If the expected legal cost in a suit that ends before summary judgment is only one-tenth of the cost of a suit that reaches summary judgment, then the estimated amount for the alleged infringer is \$211,000 and for the patentee the amount is also \$211,000.

B. *Firm Value and Patent Lawsuits*

Using our CAR estimates, we can calculate the loss of wealth that occurs upon filing a lawsuit. From this, we can then infer a cost to alleged infringers. Multiplying the estimated CAR for each firm by the value of its outstanding shares of common stock immediately prior to the lawsuit filing, we obtain a mean loss of wealth of \$83.7 million in 1992 dollars. This is an unbiased estimate of the mean loss of wealth; however, it is not the most efficient estimate. We can do better by multiplying the *mean* CAR by each firm's capitalization.⁶³

Using means for three categories—suits with multiple defendants, those with single defendants with more than 500 employees, and those with single defendants with 500 or fewer employees—we obtain a mean estimated loss of \$52.4 million and a median loss of \$4.5 million, both in 1992 dollars.⁶⁴ These estimates are somewhat smaller than Lerner's estimate for biotech companies of a mean loss of \$67.9 million and a median loss of \$20 million.⁶⁵

This loss of wealth corresponds to the associated drop in investors' expected profits. But does this loss of wealth correspond to the cost of litigation? There are two reasons why it might not. First, the filing of a lawsuit might reveal information that causes investors to revalue the firm for reasons other than the direct and indirect costs of litigation. We explore these possibilities in this section. Second, as shown in the next section, we consider how much investment the firm must undertake to restore its investors' wealth (this might not equal the loss of wealth itself).

News of a lawsuit causes investors to re-evaluate their expectations of the discounted profit flow expected from the defendant firm for several different reasons. We assume that the Efficient Market Hypothesis holds, implying that investors incorporate all publicly available information into their valuation of the firm. Consider defendant firm i at time $t = 0$, before the lawsuit filing, and at $t = 1$, immediately after the news of the filing has been made public. At $t = 0$, investors' expected value of the firm based on publicly available information, V , is

$$V_i(0) = \pi_i(0) - p_i(0)C \quad (3)$$

⁶³ The first estimator is $\frac{1}{N} \sum_{i=1}^N (r + e_i)x_i$ where N is the number of firms, r is the true CAR, e is the error in measuring the i th firm's CAR, and x is the i th firm's market capitalization. The second estimator is $\frac{1}{N} (r + \sum_{i=1}^N e_i/N) \sum_{i=1}^N x_i$. It is straightforward to show that both are unbiased but that the latter has smaller variance assuming that e and x are uncorrelated.

⁶⁴ Specifically, we multiply the common stock capitalization by .00012 for firms in cases with multiple defendants, by .00564 for single defendants with more than 500 employees, and by .0208 for small single defendants.

⁶⁵ Lerner, *supra* note 15, at 471.

where π represents the discounted expected profits of the firm (excluding litigation), p is the expected number of times the firm will be sued for patent infringement, and C is the total expected cost to the firm of a patent lawsuit. This expected cost of litigation includes:

- Legal costs.
- Indirect costs, such as management distraction, loss of market share during the lawsuit, and loss of lead-time advantage.
- Financial costs arising from greater risk, including risk of bankruptcy. These include the possibility of both higher costs of funds, and the loss of wealth associated with a higher risk-adjusted discount rate applied to the stream of future expected profits.⁶⁶
- Costs of expected outcomes including those associated with a settlement agreement and trial outcome—investors take expectations over all possible outcomes and also over the length of time and cost incurred before outcomes are reached.

It should be noted that the last term on the right hand side represents the a priori expectation of litigation cost. Then at time $t = 1$,

$$V_i(1) = \pi_i(1) - p_i(1)C - C \quad (4)$$

Comparing equation (2) and equation (1) and taking expectations over all lawsuits, the mean CAR should equal:

$$E[\Delta V] = E[\Delta \pi] - E[\Delta p]C - C \quad (5)$$

The first term on the right represents the change in investors' expectations about the future profit stream based on new information made public by the lawsuit filing. The second term in equation (5) represents investors' re-assessment of the risk of future litigation. This occurs if the lawsuit provides information that the firm is somehow more prone to litigation than originally expected. Alternatively, if investors anticipated this particular lawsuit *ex ante*, then the expectation of litigation might decrease. Clearly, if the sum of the first two terms is non-zero, then the change in firm value provides a biased estimate of the cost of litigation.

There are two sources of information from the filing that might affect these two terms:

1. Information revealed by the filing documents themselves (and any associated press releases, etc.); and,

⁶⁶ We are implying that π includes the discounted profit stream evaluated at the original discount rate. This interpretation is consistent with our definition of the cost of litigation being the level of investment necessary to restore the wealth of the firms' investors to the level just prior to the lawsuit.

2. Any information revealed by the event as a signal of the patentee's beliefs. For example, because litigation is costly, the lawsuit may signal that the patent holder believes that the opportunity at stake is particularly valuable; otherwise the suit might not be worth the cost. Note that the documents may reinforce this signal—the claim for damages may also be large, but with a signal the claim may become credible.

In order for either source to cause investors to revalue the firm, the lawsuit filing must somehow reveal information that was not previously public knowledge—under the Efficient Market Hypothesis we assume that investors correctly incorporate all public knowledge. In other words, the patent holder or the defendant firm must have some *private* knowledge that is revealed in the filing documents or by the signal generated by the filing.

Therefore, if the first two terms in equation (5) are to affect the mean CAR substantially, there must be a *systematic* reason for the patent holder or the alleged infringer to have private information that is revealed by the lawsuit filing. The documents in the lawsuit filing typically reveal relatively little hard information other than the fact of the filing, often exaggerated claims of damages, and possible allegations of bad behavior by the defendant.⁶⁷ The patents themselves, of course, are necessarily public information before the suit is filed. But we can identify three reasons why the parties might have private information that is revealed by the filing:

1. Private information about the quality of the technology. For well-known reasons, managers have private information about the quality of their technology. A lawsuit may signal that the patent holder knows that the defendant's technology is of better quality than investors previously realized, hence the market potential is greater, and a lawsuit may become more profitable. Note that in this case, $E[\Delta\pi] > 0$.
2. Private information about entry plans. If a patent holder plans on entering the defendant firm's market, then the lawsuit might reveal this knowledge, causing investors to revalue the defendant firm downwards because they expect greater competition for the firm. Note that in order for this factor to substantially affect our average CARs, such prospective entrants must initiate a substantial number of patent lawsuits. Also, the prospective entrants cannot have revealed any information about their entry plans prior to filing the lawsuit. This strikes us as a rather odd business strategy—one would think a superior strategy would be to enter the market *before* filing a lawsuit so as to capture market share from those customers who want to avoid the defendant firm. Nevertheless, we will look at

⁶⁷ See *infra* Section III.C.

empirical evidence regarding this story below. In this case, $E[\Delta\pi] < 0$.

3. Private information about managerial quality or level of effort. For well-known reasons, managers keep private information about their abilities and about the level of effort that they exert. Lawsuits might tend to indicate that managers at the defendant firm were not sufficiently diligent in clearing patent rights or, worse, that they copied technology rather than developing their own. If this tends to be true and if managers tend not to correct their behavior following a lawsuit, then investors might revalue future profits downwards. This occurs both because investors might expect more patent litigation in the future, the second term of equation (5), and because poor managerial quality might also reduce profits generally, the first term in equation (5).

However, several empirical observations lead us to discount the second and third explanations. If lawsuit filings revealed news about previously unknown entrants, we might expect these two explanations to be particularly true for plaintiffs that had recently gone public. These plaintiffs might not be widely known and therefore, on average, defendant firms might lose greater value when sued by newly public firms. However, we find that defendants' CARs are significantly more positive when the plaintiff is a newly public firm (see Table 4).⁶⁸

In addition, if news about entry is a significant factor affecting average CARs, then we would expect to find that a significant portion of plaintiffs were not known as market rivals to the defendant firm prior to the lawsuit, but rather, subsequently became market rivals. Using Compustat's market segment data, we found that this fact pattern is actually rather uncommon. Compustat reports SIC codes for each firm's major market segments. Of the plaintiffs who had no market segments in common with defendants prior to the lawsuit, we found that only 5% entered a market segment in common with the defendant during the three years following the lawsuit filing.⁶⁹ Thus, it seems unlikely that a substantial part of defendants' CARs can be explained by revelation of previously unknown entrants.

Other evidence leads us to discount the significance of any news about managerial quality or effort revealed by the lawsuit. Managerial quality is less likely to be of significance in lawsuits that are filed the same year that the patent is granted. Often these patents contain claims that were not pre-

⁶⁸ The increase could occur because startup firms are less able to pursue sustained litigation, and therefore, a lawsuit from a startup poses less of a threat. Alternatively, a lawsuit by an entrant may indicate that the technological opportunity is greater than investors previously realized.

⁶⁹ This figure compares SIC market segments at the 4-digit level. A comparable calculation using three-digit industry classifications finds a 6% entry rate. This comparison only concerns major market segments, so some entry is unrecorded in minor segments; however, rivalry in minor market segments is only likely to have a minor effect on firm value.

viously publicly known, so there is less that managers could have done to avoid infringement and managerial quality is less of an issue. For this reason, lawsuits on these patents cannot reveal as much about managerial quality. If revelations about managerial quality explain a large portion of the defendants' CARs, we would expect the CARs to be more positive for patents issued the same year as the lawsuit. In fact, we find that the CARs are more negative for these patents, although the difference is not statistically significant.

Furthermore, we would expect that the managerial quality explanation is much more significant the first time a firm is sued. That is, if a lawsuit reveals significant information about managerial quality, we would expect the second lawsuit to reveal less information, and each subsequent lawsuit to reveal even less than the one before. In particular, we would expect investors to learn and, for this reason, we would expect that, on average, CARs would reflect less revelation of information about managerial quality for, say, the fourth through tenth lawsuit than for the first three.⁷⁰ We compared defendant CARs depending on the number of lawsuits the firm had in our sample or on the sequence of the lawsuit. We found no significant differences between CARs for a wide range of different comparisons; e.g., firms with only one lawsuit in our sample had CARs that were on average only .0008 (standard deviation of .0047) less than the CARs for firms sued multiple times. Thus, revelations about managerial quality do not seem to explain much of the average loss in firm value from the filing of a lawsuit.

We have little empirical evidence bearing on the role of revelations about technological quality other than anecdote.⁷¹ In Table 4, we saw that defendants do better when the lawsuit is filed by a newly public firm. One possible explanation, though not the only one, is that suits by newly public firms reveal information about technological quality. However, as we noted above, for revelation about technological quality, $E[\Delta\pi] > 0$. Given this, we conclude that $E[\Delta\pi] \geq 0$ and $E[\Delta p] \approx 0$, so that $C \geq E[\Delta V]$. That is, the cost of litigation is likely to be at least as large as the loss in firm market value.

C. *Investment Level Costs*

If we want to know how much litigation “taxes” investment in innovation, then we need to calculate something other than the loss of wealth. That is, all else equal, we define the “cost of litigation” as the amount that the firm has to invest in order to increase its value to the level it had just prior to the lawsuit. This does not necessarily equal the amount of wealth

⁷⁰ This assumes, of course, that management is not entirely replaced between lawsuits.

⁷¹ A tech industry joke on hearing that someone has been sued is: “Congratulations, you must be doing something right!”

the firm loses because firms are not necessarily operating at the long-run steady state. Instead, they may be undergoing dynamic adjustment. Therefore, changes in investment will be larger or smaller than the associated changes in firm value. In particular, assuming constant returns to scale, an additional investment of one dollar should increase firm value by an amount equal to Tobin's Q .

Following this logic, in order to calculate the cost of litigation, we divide the estimated loss of wealth by Tobin's Q .⁷² This gives us a mean cost of litigation to alleged infringers of \$28.7 million and a median cost of \$2.9 million in 1992 dollars.

These estimates are clearly much larger than the estimates of direct legal costs. Most of the cost of litigation to firms appears to arise from expected settlement payments and business costs such as loss of market share, management distraction, and increased financial costs from greater risk. These costs are incurred even if the suit does not proceed to trial, as happens most often.

It is interesting to compare our estimate to data from cases that proceed to trial. For the small number of reported cases that go to trial, are won by the patentee, and in which damages are awarded to the patentee, we can compare the magnitude of these damages. Mean reported lawsuit damages from 1991 to 2005 were \$10.7 million in 1992 dollars.⁷³ This number does not include the business cost of the injunction to the infringer—often much larger than the damages. For example, the court found damages of \$53.7 million in *NTP v. RIM*,⁷⁴ but because of the injunction, NTP eventually settled for \$612 million.⁷⁵ This mean also does not include the costs of pursuing the litigation, both direct payment of legal costs, and indirect business costs. Nevertheless, it is reassuring that this figure is of the same order of magnitude as our mean estimate.

IV. THE RISK OF INFRINGEMENT FOR PUBLIC FIRMS

These cost estimates can be summed over all the observed lawsuit filings to obtain measures of firm risk. Table 7 shows three related measures.

⁷² See James Bessen, *Estimates of Firms' Patent Rents from Firm Market Value* 3 (Boston Univ. Sch. of Law, Working Paper No. 06-14, 2006). We calculate Tobin's Q as the aggregate value of firms divided by the inflation-adjusted value of the aggregate sum of accounting assets and R&D.

⁷³ See PRICE WATERHOUSE COOPERS ADVISORY CRISIS MGMT., 2006 PATENT AND TRADEMARK DAMAGES STUDY 11 (2006). This figure is the mean of deflated annual means.

⁷⁴ *NTP, Inc. v. Research In Motion, Ltd.*, 418 F.3d 1282, 1292 (Fed. Cir. 2005).

⁷⁵ Christopher Rhoads, *Mixed Messages: In BlackBerry Case, Big Winner Faces His Own Accusers --- Stout Received \$177 Million But Some Ask Why Firm He Leads Got Key Patents—A Scorned Creditor's Fury*, WALL ST. J., Aug. 23, 2006, A1.

Table 7. Measures of Infringement Risk, Public Firms

	Aggregate Annual Cost of Litigation to Alleged Infringers (billion \$92)	Annual Firm Infringement Risk (million \$92)	Aggregate Risk/R&D
1984	2.0	1.3	4.9%
1999	16.1	7.0	19.3%
<u>1996–99</u>			
All firms	14.9	4.5	14.0%
Small firms (employees <500)	0.1	0.1	1.3%
Large firms (employees > =500)	14.8	9.8	14.9%
SIC = 28 (chemicals, inc. pharma)	3.4	9.7	14.1%
SIC = 35, 36, 73 (electronics, computer, software)	6.8	5.7	14.8%
Other manufacturing	1.7	2.3	5.3%

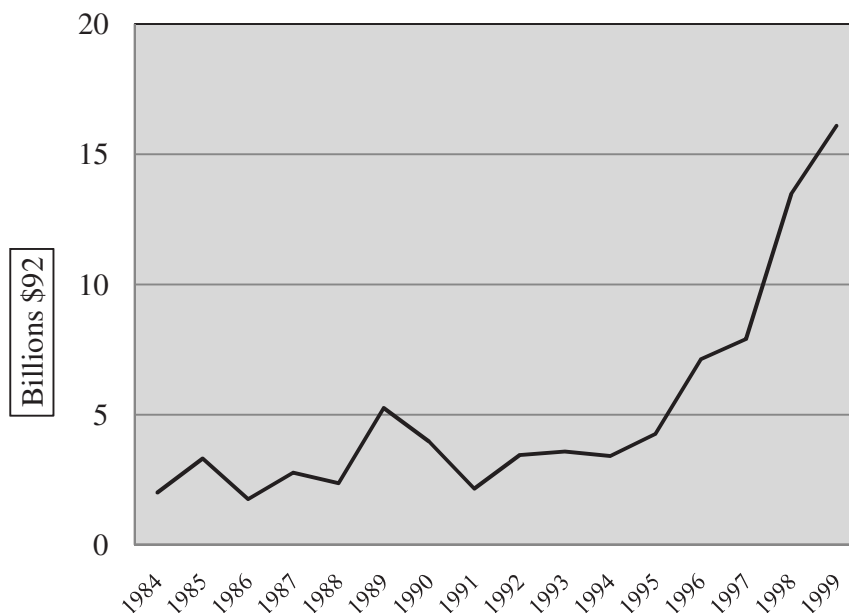
Note: The annual cost of litigation is the mean CAR times the market capitalization of each firm's common stock divided by a GDP deflator and by the aggregate Tobin's Q (market value divided by replacement value of capital including R&D). Firm infringement risk is the expected annual cost of litigation. Column 1 includes all events in the large sample (2,887) with separate means for small firms and lawsuits with multiple defendants. Columns 2 and 3 have been adjusted for under-reporting of lawsuits. See Lanjouw & Schankerman, *supra* note 17; Bessen & Meurer, *supra* note 10.

The first column lists the annual cost of litigation obtained by summing the cost over all the events in our large sample in each year of the sample. During 1996 to 1999, this averaged \$14.9 billion in 1992 dollars. This number is large compared to estimates of patent value. Using renewal data to estimate patent value, Bessen, reports the aggregate value of patents

issued to *all* U.S. patentees, not just public firms, in 1991 was about \$4.4 billion.⁷⁶

Moreover, this figure has varied considerably over time, increasing dramatically from \$2 billion in 1984 to \$16.1 billion in 1999. Figure 3 shows the annual time series. The rise began in the early 1990s and closely follows the increasing frequency of litigation.⁷⁷ Other factors contributed as well, including the increase in R&D spending and firm capitalization. Below, we look at infringement risk normalized by R&D. The absolute cost of litigation was borne almost entirely by large firms and nearly half by firms in the computer, electronics, and software industries.

Figure 3. Aggregate annual cost of patent litigation to alleged infringers



Note that this series may be substantially understated because, as is well-known, the Derwent Litalert data under-report lawsuits.⁷⁸ In our 2005 working paper using this sample, we find that only about 64% of lawsuits are reported in Derwent.⁷⁹ We have left the first column of Table 7 uncorrected, since it reports a simple sum for our sample. However, the second and third columns compare litigation cost to numbers of firms and to R&D

⁷⁶ James Bessen, *The Value of U.S. Patents by Owner and Patent Characteristics* (Boston Univ. Sch. of Law and Econ., Working Paper No. 37, 2008).

⁷⁷ Bessen et al., *supra* note 8, at 2-3.

⁷⁸ *See id.* at 11-12; Lanjouw & Schankerman, *supra* note 17, at 49-50 (2004).

⁷⁹ Bessen et al., *supra* note 8, at 12.

spending, respectively. In order to make the appropriate comparisons, we correct these for under-reporting by dividing by 0.64.⁸⁰

On the other hand, this series may slightly overstate the aggregate cost of patent litigation per se because some of the suits listed involved more than just charges of patent infringement and validity. For example, sometimes patent owners will combine allegations of patent infringement with allegations that other rights, including other intellectual property rights, have been violated. Some of the suits of this sort might occur even if patent infringement was not at issue, so it might not be appropriate to include all of the costs associated with these suits in an aggregate estimate of patent litigation costs. However, we do not think this is a serious problem for two reasons. First, from a search of published court decisions between 1991 and 1999, only 11% of patent infringement and validity suits also involved claims involving trade secrets, trademarks, copyright, false advertising, unfair competition, or noncompete clauses.⁸¹ Second, in Table 4 we observed that the alleged infringer's losses are much greater for inter-industry suits than for intra-industry suits. Since most of the cases involving these additional legal issues occur between rivals in the same industries, these suits do not contribute much to aggregate litigation costs. Accordingly, it seems unlikely that our aggregate cost estimates overstate the costs of patent litigation by more than a few percent.

The second column of Table 7 displays the annual firm infringement risk. This is the mean expected cost of litigation for a firm from patent infringement lawsuits or related declaratory actions. It averaged \$4.5 million between 1996 and 1999, and it shows a similar pattern of distribution.

The third column of Table 7 displays the ratio of annual litigation cost to annual aggregate R&D. This averaged 14.0% between 1996 and 1999. This relative rate also increased from 1984 to 1999, more than tripling to 19.3%—roughly in line with the growth of the litigation hazard. However, this increase was not as rapid as for the quantity in column 1. Note that relative to R&D, litigation risk is low for small firms and for firms outside of the chemical, pharmaceutical, and tech industries.

It is tempting to compare this ratio with the “equivalent subsidy rate” for patents—the aggregate value of patents divided by the value of the corresponding R&D. Schankerman suggests that this ratio represents an upper bound on the subsidy that patents provide to invest in innovation.⁸² But, as we argued above, this is clearly a gross subsidy that can be offset by litigation risk, if innovators risk inadvertent infringement, and by other costs. Several papers calculate this ratio by comparing the value of a nation's patents, estimated using patent renewal data, to R&D, calculated by allocating

⁸⁰ Lanjouw & Schankerman, *supra* note 17, at 49-50 (finding no significant differences between the characteristics of the reported and unreported lawsuits).

⁸¹ Based on a search of case synopses in the Westlaw FIP-CS database.

⁸² Schankerman, *supra* note 4.

national R&D spending to the patents obtained in the subject country. Lanjouw et al. reviewed this literature and reported that most subsidy rates are on the order of 10–15%.⁸³ Arora et al. used survey data to obtain a comparable estimate of 17%.⁸⁴

However, these numbers are not directly comparable to our estimates of relative litigation risk for at least three reasons. First, because of the way these studies allocate global R&D, they effectively report the subsidy provided by worldwide patents, not patents in a single country.⁸⁵ However, the litigation cost is only for U.S. litigation and does not include the costs of litigation in other countries. Nor does it include the costs of other dispute resolutions such as opposition proceedings. An “apples-to-apples” comparison would include these costs as well.

Second, the subsidy rate calculations based on patent value use the value of all of the nation’s patents, including patents from individual inventors and small firms. The litigation risk estimates are only for public firms—the firms that conduct the lion’s share of R&D. A more appropriate comparison would be to calculate subsidy rates using patent values only for public firms.⁸⁶ In any case, public firms may experience both different subsidy rates and different litigation costs than other firms.

Finally, the litigation costs are estimated for the current year, but the value of patents granted reflects a stream of profits in *future* years. Ideally, we would want to compare litigation costs to the profits from patents on the same cohort of technologies that were litigated. Some of these profits are realized prior to the time of litigation. Since both litigation costs and patent values are trending up, this use of current patent values understates the significance of litigation costs.

All three of these considerations suggest that a direct comparison of reported subsidy rates to US litigation risk overstates the relative positive value of patents. At the very least, these estimates suggest that litigation risk is quite large compared to the private benefits of patents, especially in recent years.

⁸³ Jean O. Lanjouw, Ariel Pakes & Jonathan Putnam, *How to Count Patents and Value Intellectual Property: The Uses of Patent Renewal and Application Data*, 46 J. INDUS. ECON. 405, 424 (1998).

⁸⁴ Arora et al., *supra* note 3, at 32.

⁸⁵ That is, using trade data, they allocate a share of the R&D performed in every OECD country to, French patents, for example, when they calculate the subsidy rate using the value of French patents. The apparent assumption behind this allocation is that subsidy rates are the same across nations and that the share of trade is proportional to each nation’s share of worldwide patent value. As such, the calculated subsidy rate will represent the return from worldwide patents. See Arora et al., *supra* note 3 (similarly using U.S. patents as a right hand variable, but note that this serves as a proxy for each firm’s worldwide patents).

⁸⁶ See James Bessen, *The Value of U.S. Patents by Owner and Patent Characteristics* (Boston Univ. Sch. of Law and Econ., Working Paper No. 37, 2008), available at http://papers.ssm.com/sol3/papers.cfm?abstract_id=949778 (showing comparable figures).

CONCLUSION

Using a large set of event studies, we estimate the total cost that patent litigation imposes on firms and we estimate the risk of infringement litigation. We find that, contrary to what is sometimes assumed, the business costs of litigation far exceed the direct legal costs. And we find that by the late 1990s, patent litigation risk was of the same order as, if not larger than, estimates of the private benefits firms receive from patents. Moreover, consistent with the previous literature, the losses to alleged infringers do not correspond to a transfer of wealth to patent holders; instead there is a substantial joint loss of wealth. Our estimates concern private costs rather than the social costs of litigation, nevertheless these estimates tell us something about the effectiveness of patents as a policy tool to encourage investment in innovation.

In the best case, this suggests that the patent system is at present an inefficient form of subsidy or regulation. Thomas Hopkins estimates the total 1992 cost of general regulatory compliance is \$389,911 per firm (in 1995 dollars).⁸⁷ But the costs of complying with the patent system—with an annual infringement risk of \$4.5 million—are much larger.

In the worst case, the net effect of patents today may be to reduce the profits of public firms and to possibly impose disincentives on innovation as well. More extensive exploration of the possible causes and their significance of this for policy and for normative analysis are beyond the scope of this paper, however. Nevertheless, our analysis indicates that infringement risk should be an important critical consideration in the formulation of patent policy.

APPENDIX

This appendix further explores our choice of a window around the lawsuit filing date rather than an announcement in a newspaper or wire service. First, we explore whether a sample based on Wall Street Journal articles is likely to suffer sample selection bias. Table A1 shows Probit regressions on whether a lawsuit in our matched sample received mention in The Wall Street Journal. The patentee litigant's capital intensity and the alleged infringer's stock beta are both highly significant predictors, at the 1% level, of a Wall Street Journal article. Because high beta stocks are likely to have a larger reaction to news of a lawsuit, this suggests that samples based on Wall Street Journal articles may have significant bias. We

⁸⁷ OFFICE OF THE CHIEF COUNSEL FOR ADVOCACY, U.S. SMALL BUS. ADMIN., THE CHANGING BURDEN OF REGULATION, PAPERWORK, AND TAX COMPLIANCE ON SMALL BUSINESS (1995), available at http://archive.sba.gov/advo/laws/archive/law_brd.html.

found, in fact, the estimates from our sub-sample of lawsuits announced in The Wall Street Journal do have much more negative CARs.

Table A1. Suit Announcement and Type

	Wall Street Journal Article		Infringement Suit	
	1	2	3	4
<u>Plaintiff/patentee litigant</u>				
Ln employment		0.05(.03)	.02(.03)	.01 (.04)
New firm		-.25(.23)	.63 (.29)	.62 (.33)
Stock Beta	.13(.12)	.15(.11)		.20 (.13)
Capital/employee	1.01 (.38)	1.12 (.40)		-.64 (.49)
<u>Defendant/alleged infringer</u>				
Ln employment		-.01(.03)	.06(.03)	.07 (.03)
New firm		.28(.20)	-.01(.20)	-.05 (.22)
Stock Beta	.35 (.13)	.35 (.13)		.05 (.14)
Capital/employee	.05(.36)	.11(.36)		-.95 (.51)
No. of observations	637	637	507	475
Pseudo-R-squared	.049	.062	.023	.057

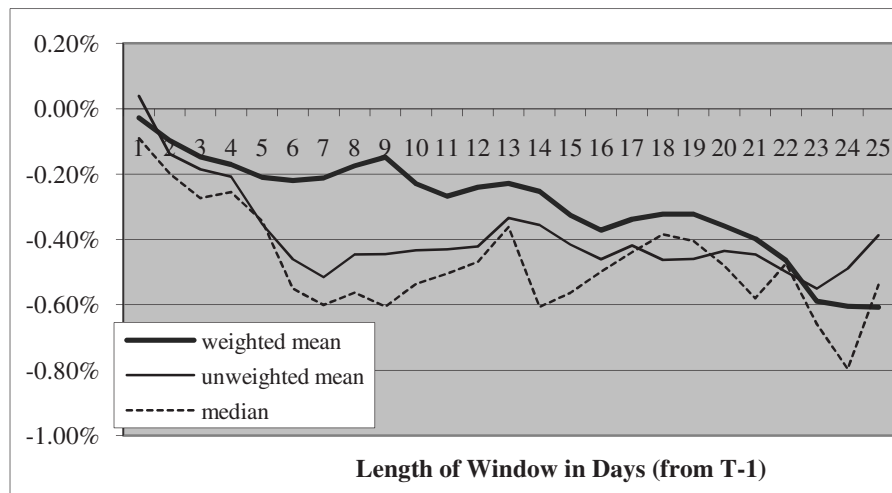
Note: Probit regressions. Robust standard errors in parentheses. Bold estimates are significant at the 5% level or better. Regressions include industry dummies (not shown).

We do not have information on whether a suit is an infringement suit or a declaratory action in all cases. Because of this, we likely misidentify some plaintiffs and defendants, resulting in the dilution of our estimates for alleged infringers. One way to correct for this would be to limit our sample to cases of definite infringement, although this may introduce a selection bias. The last two columns of Table A1 explore characteristics that may affect whether the suit is an infringement suit or a declaratory action. It appears that newly public patentees may be slightly more aggressive in filing suits, while larger alleged infringers may be more likely to end up in an infringement suit. Large firms may avoid filing declaratory actions because they are waiting for evidence that the patent owner has the resources to conduct a lawsuit. We report CARs both for the entire sample and for

cases that we know are infringement suits to take into account the possibility of the existence of a selection bias.

Finally, as discussed in the text, because news of a lawsuit filing leaks out more slowly than a newspaper announcement, we use a twenty-five day event window. Figure A1 shows the mean CARs we would obtain using shorter event windows. Note that the unweighted mean and median CARs both react more sharply in the days after the filing. This is because high beta stocks respond more quickly after the filing—they are the ones where investors may have greater incentive to obtain such news. Because the CARs for low beta stocks are estimated more precisely and their response is slower, the weighted mean responds more slowly. However, all three averages are roughly equal by the end of our twenty-five day window.

Figure A1. Average Abnormal Cumulative Returns Over Time





NEXT-GENERATION COMPETITION: NEW CONCEPTS FOR
UNDERSTANDING HOW INNOVATION SHAPES COMPETITION AND
POLICY IN THE DIGITAL ECONOMY

*David J. Teece**

INTRODUCTION

In advanced economies with good infrastructure, good public policies, rule of law, and strong property rights, innovation is the primary driver of economic growth. This means not only innovation in goods and services but also innovation in the way that businesses operate, both individually and in combination. Though often overlooked, organizational and managerial innovations are as important to economic growth as technological innovation.

Technological innovation may have accelerated in recent decades and is unquestionably shaping the competitive landscape. Moreover, with the advent of the Internet, the impact of organizational innovation is more salient as new business models enhance performance and permit the viability of new types of businesses.

One can better distill policy and management implications by observing the changes that are afoot. Perhaps unsurprisingly, our present understanding of these unfolding developments is rudimentary. Academics are having only modest levels of success in comprehending, interpreting, and informing the evolution of this new order while practitioners are dealing with the new phenomena on a daily basis.

Businesses must react in real time to changes in technology, regulation, and competition. The most alert and agile leadership teams are out in front driving innovation and change. Business and legal scholars closely follow evolving business practices and observe them first-hand. They must evolve their frameworks to keep up with changes as best they can in order to stay relevant. Many are endeavoring to do so.

Unfortunately, mainstream economists are slow in coming to understand this new order. In the last half of the 20th century, the economics profession collectively cloaked itself in the standards of “good science.” Economists, unfortunately, seem to have interpreted good science as requiring the use of the models and theories of neoclassical economics. These models and theories, which often suppress institutional, managerial, and

* Director, Institute for Business Innovation, Haas School, UC Berkeley. Paper based on luncheon address at The Digital Inventor: How Entrepreneurs Compete on Platforms, George Mason University, on February 24, 2012. I wish to thank Greg Linden for helpful comments and assistance.

technological factors, obscure more than they reveal. A few bright spots aside, the mathematical models respected in the discipline of economics are woefully behind, and mostly caricature real-world developments. The elegance of mathematical models has become the paramount criterion for judging good research, often to the point that institutional and business realities are ignored, which leaves the models deeply flawed.

Lastly, there are the policy analysts. Their comprehension is retarded by a combination of politics, bureaucracy, confusion about what to accept from the academic disciplines, and a bias toward holding on to familiar analytical frameworks.

With these differences in mind, this essay will review some of the most exciting recent developments in the way that businesses operate, collaborate, and compete. Each of these features of the competition and innovation landscape will be well-known to some or all who participate in the technology industries. Nevertheless, there is value in discussing and reframing these new ideas, as they have yet to fully permeate the academic and policy literature.

As noted earlier, new technologies and practices are sharpening competition; competition is, in turn, reshaping business institutions and practices—and so on. This new turbo-charged competition is so different from the scale-based competition of the previous century that it deserves to be called next-generation competition. Next-generation competition is changing the way businesses compete, collaborate, and operate.

It is useful to contrast these new concepts with the established ideas that still frame much mainstream analysis. The concepts that this essay will discuss in some detail are shown in Table 1. This list by no means exhausts the next-generation phenomena occurring in the business world, but it does give some flavor of them.

Table 1. Old and New Modes of Competition

Conventional Concept	Next-Generation Concept
Static Competition	Dynamic Competition
The West and the Rest	A Semi-Globalized World
Industry-level Analysis	Ecosystem-level Analysis
Vertical Integration	Modularization
Transaction and Agency Costs	Firm-level Capabilities
Single-Invention Innovation Model	Multi-Invention Innovation Model

I. DYNAMIC COMPETITION

In standard formulations, such as industrial organization economics, or Michael Porter's "Five Forces,"¹ competition is determined primarily by market structure: concentrated—possibly "monopoly"—market structures result in high prices; oligopoly market structures result in indeterminate prices; and perfect competition market structures result in prices that are low and precisely equal to marginal cost. Market structure, in turn, shapes innovation. In some formulations, the greater the market power, the greater the rents available for supporting research and development (R&D).

However, in the rapidly changing real world, incumbent firms can seldom gain durable advantage from high market shares. Start-ups and firms from related industries move in quickly to create new rent streams that undermine existing business models. In 2009, RIM, the developer of BlackBerry mobile devices, was the second-most profitable company in the cell phone business. Today, it is an open question whether RIM can survive as an independent company because of the challenges posed by Apple and Google—the latter being a company that does not even sell a purely own-brand phone. As this example suggests, market structure can be shaped not only by technological innovation, but by business model innovation as well.

In other words, causation runs in the opposite direction from what is commonly assumed in standard textbook treatments of the competition-

¹ See MICHAEL E. PORTER, *COMPETITIVE STRATEGY* 3-7 (1980).

innovation nexus. In fact, new entrants have been responsible for a substantial share of revolutionary new products and processes for a long time. These include the jet engine (Whittle in England; Henkel and Junkers in Germany), catalytic cracking in petroleum refining (Houdry), the electric typewriter (IBM), electronic computing (IBM), electrostatic copying (Haloid), PTFE vascular grafts (WL Gore), the microwave oven (Raytheon), diet cola (RC Cola), and wireless handsets (Apple). These anecdotes and other evidence further indicate that there is a lack of any meaningful causal connection between market power and innovation.² Yet static analysis still permeates much of economic theory. Economists are too enamored with models that yield an equilibrium, while the world they need to explain is in perpetual disequilibrium.

For policymakers, the Horizontal Merger Guidelines,³ which serve as an intellectual cornerstone of modern antitrust law, contain little discussion of innovation, focusing on the technology trajectories of the merging firms while ignoring the activities of existing and emerging rivals unless their entry is “timely, likely, and sufficient.”⁴ Some scholars have recognized the importance of dynamic competition for decades.⁵ However, only recently have more mainstream antitrust scholars begun to actively debate the merits of replacing static competition with dynamic competition in antitrust analysis.⁶ It remains to be seen when the governing logic of competition policy

² See David J. Teece, *Profiting from Technological Innovation*, 15 RESEARCH POL’Y 285 (1986); see also DAVID J. TEECE, MANAGING INTELLECTUAL CAPITAL: ORGANIZATIONAL, STRATEGIC, AND POLICY DIMENSIONS 39-43 (2000) (providing frameworks and advice on several aspects of innovation).

³ U.S. DEP’T OF JUSTICE & FED. TRADE COMM’N HORIZONTAL MERGER GUIDELINES (2010), available at <http://www.justice.gov/atr/public/guidelines/hmg-2010.html>.

⁴ *Id.* at § 9.

⁵ See generally Thomas M. Jorde & David J. Teece, *Antitrust Policy and Innovation: Taking Account of Performance Competition and Competitor Cooperation*, 147 J. INST. AND THEORETICAL ECON. 118, 119-121 (1991); J. Gregory Sidak & David J. Teece, *Dynamic Competition in Antitrust Law*, 5 J. COMPETITION L. & ECON. 581, 583 (2009); David J. Teece & Mary Coleman, *The Meaning of Monopoly: Antitrust Analysis in High-Technology Industries*, 43 ANTITRUST BULL. 801 (1998).

⁶ See, e.g., Jonathan Baker, *Dynamic Competition’ Does Not Excuse Monopolization*, 4 COMPETITION POL’Y INT’L 243 (2008); Christian Ewald, *Competition and Innovation: Dangerous ‘Myopia’ of Economists in Antitrust?*, 4 COMPETITION POL’Y INT’L 253 (2008); Richard Gilbert, *Injecting Innovation into The Rule of Reason: A Comment on Evans and Hylton*, 4 COMPETITION POL’Y INT’L 263 (2008); Herbert Hovenkamp, *Schumpeterian Competition and Antitrust*, 4 COMPETITION POL’Y INT’L 273 (2008); Michael L. Katz & Howard A. Shelanski, *Schumpeterian Competition and Antitrust Policy in High-Tech Markets*, 14 COMPETITION 47 (2005), available at www.law.berkeley.edu/institutes/bclt/pubs/shelanski/katz_Shelanski_Schumpeter__30Nov2006_final.pdf; see also Thomas K. McCraw, *Joseph Schumpeter on Competition*, 4 COMPETITION POL’Y INT’L 309 (2008); Richard Schmalensee, *Standard-Setting, Innovation Specialists, and Competition Policy*, 57 J. INDUS. ECON. 526 (2009) (discussing the role of dynamic competition in the antitrust analysis of patent royalties and standard-setting); Ilya Segal & Michael D. Whinston, *Antitrust in Innovative Industries*, 97 AM. ECON. REV. 1703 (2007).

will fully recognize the new reality. In the meantime, policy errors, in the form of unnecessary interventions, are likely to continue.

II. A SEMI-GLOBALIZED WORLD

The West has long been—and, many think, still is—seen as pre-eminent in technology and business. This is fast becoming a dangerous illusion as other countries replicate—and, sometimes, improve upon—the key technologies and management systems that were invented in the West.

The globalization (and quasi-virtualization) of higher education over the past twenty years has provided support for the efforts of companies in industrializing countries to develop competitive R&D capabilities. The growing international dispersion of applied R&D results in many sources of innovation, and hence potential innovation partners. Global and organizational dispersion in the sources of know-how brings the need, and the opportunity, for open innovation. Open innovation employs new mental models and processes to tap quickly, efficiently, and effectively into the great diversity of ideas, know-how, and solutions to theoretical and practical problems.⁷

The global dispersion and diversity of knowledge is joining with the dispersion of (organizational) capabilities to make a much more complex and competitive landscape. As firms from less-developed countries begin to innovate by solving the problems faced by their local customers, they are increasingly able to combine this with marketing and operational know-how to build an advantage over the major multinationals in other developing markets.

The global dispersion of knowledge, combined with low-friction global transportation and information flows, has caused some to say that the world is “flat.”⁸ While it is more correct to see the world as semi-globalized,⁹ it is nevertheless true that intermediate goods and services that might once have been difficult to access are now widely available—a reali-

⁷ See David J. Teece, *Technology and R&D Activities of Multinational Firms: Some Theory and Evidence*, in TECHNOLOGY TRANSFER AND ECONOMIC DEVELOPMENT 369 (R. G. Hawkins & A. J. Prasad eds. JAI Press 1981). See generally HENRY W. CHESBROUGH, OPEN INNOVATION: THE NEW IMPERATIVE FOR CREATING AND PROFITING FROM TECHNOLOGY 43-62 (2003); Gary Pisano & David J. Teece, *Collaborative Arrangements and Global Technology Strategy: Some Evidence from the Telecommunications Equipment Industry*, in RESEARCH ON TECHNOLOGICAL INNOVATION, MANAGEMENT AND POLICY 4 (1989).

⁸ See generally THOMAS L. FRIEDMAN, THE WORLD IS FLAT: A BRIEF HISTORY OF THE TWENTY-FIRST CENTURY (2007).

⁹ See Pankaj Ghemawat, *Semiglobalization and International Business Strategy*, 34 J. INT'L BUS. STUDIES 138, 139 (2003); see also Pankaj Ghemawat and Fariborz Ghadar, *Global Integration ≠ Global Concentration*, 15 INDUS. & CORP. CHANGE 595 (2006) (indicating that increasing global integration has not eliminated all local market distinctions).

ty which has created a system of globally-distributed specialization. On this not quite “flat,” but, rather, gently undulating landscape, the capabilities required to orchestrate and deploy the available resources remain scarce and geographically isolated. The traditional competitive sources of differentiation based on economies of scale and scope have been eroded because almost everything can now be outsourced. If a firm’s target market is too small to let it capture economies of scale for an input or even a whole product, then it should source from companies that have achieved them already. Some textbooks still need to be rewritten to recognize this new reality.

Fortunately for business firms and their stakeholders, there are certain other residual bases of competitive advantage unrelated to scale. These are even more salient, given that the more traditional ways of generating points of difference have eroded. The primary way that differentiation can be created and sustained is through the generation, ownership, and management of intangible assets. Intangibles have risen to overshadow economies of scale and scope in their importance for enabling the enterprise to build and sustain competitive advantage.

Even in natural resource industries, a deeper inquiry will almost always reveal that profits flow—perhaps surprisingly—from the ownership and use of intangibles, and less so from the natural resource itself. Profits flow to those who develop safe, efficient, and effective extraction technologies or who build privileged relationships with nation-states and owners of natural resources so as to obtain exploration and extraction rights on favorable terms. For example, speaking metaphorically, oil is “found” in the mind—by employing knowledge assets—not in the ground; locating and bringing to market new crude oil reserves requires both (organizationally embedded) know-how and, in many jurisdictions, relationships with nation-states. Both classes of intangible assets, know-how and relationships, are the keys to finding and producing crude oil profitably. The paradox here is that intangible assets are of supreme importance even in the natural resource or extractive industries.

If true in so-called natural resource industries, it is unquestionably true in the so-called “tech sector” that intangibles are of supreme importance. For start-up firms, this requires management to identify the intangibles that make them unique, namely the expertise, know-how, patents, etc., that will be difficult to replicate. Of course, success also requires that such intangibles are relevant to addressing a large and growing market. Firms also need to scan the world for the partnering and marketing opportunities that make the most sense for their stakeholders. A growing number of start-ups are global from birth, with engineering on one continent, manufacturing in another, and markets elsewhere.

Economists and other social scientists are still a long way from coming to grips with the significance of intangible assets. At present, the social sciences might grudgingly acknowledge the importance of such assets, but the analysis then proceeds to ignore them. Yet intangible assets, of which a

firm's knowledge and relationship assets constitute the most important classes, are at the heart of enterprise profitability.¹⁰ Such assets are hard to "build" and difficult to manage. They are also unlikely to be traded—markets, if they exist, will be "thin"—because their underlying value derives from the presence of complementary assets, which limits the number of buyers who will be willing and able to pay the knowledge asset's full potential strategic value.

Knowledge assets, which are tacit to varying degrees, are generally costly to transfer and can even be difficult to specify fully in a contract.¹¹ As a result, knowledge assets are harder to access than many other asset types. This confers special competitive status. Firms that strategically create, deploy, and protect them have a chance to build a durable competitive advantage.

Prices of knowledge assets are generally unobservable. This is a corollary of the fact that markets for these assets are thin. Financial analysts might say that there is almost no liquidity in the market for know-how and for intangible assets more generally. Without going to extraordinary efforts, the value of such assets can at best only be estimated. Moreover, the process by which these assets are generated is virtually unmeasurable as to inputs and, possibly, also outputs. This makes these assets difficult for economists to model and even harder for policymakers to encompass in their thinking. They are absolutely critical to building enterprise-level competitive advantage, yet the textbooks barely mention them. Industrial policy everywhere struggles to deal with this new reality. Economists and policy makers recognize the role of intangibles but rarely explore it.

Politics in the United States are currently hostile to affirmative and coordinated innovation policies, settling for a patchwork of programs and policies that often hurt as much as they help. Other nations more steadfastly maneuver to favor domestic industry based on a better appreciation of organizational learning and capability development. For instance, U.S. antitrust policy effectively swept away Bell Labs through the breakup of AT&T—inadvertent collateral damage perhaps, yet entirely predictable and extremely harmful to the long-term health of the U.S. system of innovation. Meanwhile, China and the E.U. are forthright about using competition policy to harm foreign competitors. For example, in March 2011, China's Ministry of Commerce delayed a billion-dollar acquisition by Nokia Siemens Networks of a Motorola line of business for sixty days of further review,

¹⁰ See DAVID J. TEECE, *DYNAMIC CAPABILITIES AND STRATEGIC MANAGEMENT: ORGANIZING FOR INNOVATION AND GROWTH* 196-97 (2009).

¹¹ See, e.g., David J. Teece, *The Market for Know-How and the Efficient International Transfer of Technology*, 458 ANNALS AM. ACAD. POL. & SOC. SCI. 81, 83 (1981).

apparently to help bring about the resolution of an intellectual property dispute between Motorola and China's Huawei.¹²

III. BUSINESS ECOSYSTEMS

Economic models and regulatory frameworks typically concentrate on industries or "markets." While theoretically tractable, these constructs correspond less and less to the reality of innovation-driven competition. They also fail to take account of another important phenomenon—business ecosystems. A business ecosystem contains a number of firms and other institutions that work together to create and sustain new markets and new products. The co-evolution of the system is typically reliant on the technological leadership of one or two firms that provide a platform around which other system members, providing inputs and complementary goods, align their investments and strategies.

An ecosystem may be anchored by a platform. A platform exists when the elements of the ecosystem depend upon common standards and interfaces.¹³ Platforms are usually proprietary in that the standards are protected by patents or copyright. Platforms typically result in specialization by ecosystem members, resulting in shorter development times for new-generation products and services. The viability of any business ecosystem depends on a platform innovator cooperating with the providers of complements and vice versa. Participants—or "members"—in the ecosystem collectively jockey for position against rival ecosystems, as in the case of the two personal computer ecosystems, one based on the Windows operating system and the other on the Macintosh.

Business ecosystems are relatively new. The world of mass production that Alfred Chandler described exhibited deep vertical integration.¹⁴ Ecosystem development was not center stage because most aspects of the value chain were under the control of a single enterprise. Put differently, with the Chandlerian corporation, the ecosystem was internalized.

Evolutionary economists such as Richard Nelson and Sidney Winter have long championed the application of evolutionary theorizing to eco-

¹² Leena Rao, *Nokia Siemens Closes \$975M Acquisition Of Motorola Solutions' Wireless Network Assets*, TECHCRUNCH (Apr. 29, 2011), <http://techcrunch.com/2011/04/29/nokia-siemens-closes-975m-acquisition-of-motorola-solutions-wireless-network-assets/>.

¹³ David Robertson & Karl Ulrich, *Planning for Product Platforms*, 39 SLOAN MGMT. REV. 19, 20 (1998).

¹⁴ ALFRED D. CHANDLER, JR., *STRATEGY AND STRUCTURE: CHAPTERS IN THE HISTORY OF THE INDUSTRIAL ENTERPRISE* 387 (1962); ALFRED D. CHANDLER, JR., *THE VISIBLE HAND* 285-86 (1977); ALFRED D. CHANDLER, JR., *SCALE AND SCOPE: THE DYNAMICS OF INDUSTRIAL CAPITALISM* (1990); see also *MANAGEMENT INNOVATION: ESSAYS IN THE SPIRIT OF ALFRED D. CHANDLER, JR.* (2012).

nomics and to the study of organizations.¹⁵ The emergence of group-based competition began to receive scholarly attention as early as the 1980s.¹⁶ The ecosystem metaphor, popularized by James Moore,¹⁷ combines these concepts to encompass the process by which entities (be they species or organizations) become enmeshed in an ongoing cycle of interdependent changes such as platform-based competition.¹⁸

A business ecosystem is typically created by an innovator choosing which elements of the value chain must be internalized, and deciding what needs to be supported externally, in order to provide it the best opportunity for capturing value.¹⁹ Co-evolution, in which the attributes of two or more organizations become more closely complementary, and co-creation, in which two or more organizations combine forces to pioneer new markets, are two key characteristics of a business ecosystem.²⁰

Bill Gates, co-founder of Microsoft, constructed such an ecosystem around Microsoft's Windows operating system. Under Gates's leadership, Microsoft reached out to application developers, PC makers, chip makers and users. In a 2002 message to Microsoft managers, Gates noted the following:

A product with high share generates a common sense around it. A common sense that Community Colleges train on that product. A common sense that temporary workers know the product. A common sense that certification in the product is a valuable thing. A common sense that the industry can exchange data or aggregate data using schema specific to that product. A common sense that someone doing something new should move to that product. A common sense in terms of how the press covers the product and its development.²¹

However, Microsoft Windows is also an example of how ecosystems can sometimes be poorly managed. Over time, Microsoft often saw com-

¹⁵ RICHARD R. NELSON & SIDNEY G. WINTER, *AN EVOLUTIONARY THEORY OF ECONOMIC CHANGE* (1982).

¹⁶ See, e.g., J. Carlos Jarillo, *On Strategic Networks*, 9 STRATEGIC MGMT. J. 31, 41 (1988).

¹⁷ James F. Moore, *Predators and Prey: A New Ecology of Competition*, HARV. BUS. REV., May-June 1993, at 75.

¹⁸ There is also a large literature on organizational ecology in the organizational behavior field. See MICHAEL T. HANNAN & GLENN R. CARROLL, *DYNAMICS OF ORGANIZATIONAL POPULATIONS: DENSITY, AND COMPETITION* (1992); MICHAEL T. HANNAN & JOHN H. FREEMAN, *ORGANIZATIONAL ECOLOGY* (1989); MICHAEL T. HANNAN & GLENN R. CARROLL, *DYNAMICS OF ORGANIZATIONAL POPULATIONS: DENSITY, AND COMPETITION* (1992).

¹⁹ Teece, *Profiting from Technological Innovation*, *supra* note 2.

²⁰ Christos N. Pitelis and David J. Teece, *Cross-border Market Co-Creation, Dynamic Capabilities and the Entrepreneurial Theory of the Multinational Enterprise*, 19 INDUS. & CORP. CHANGE 1247, 1270 (2010).

²¹ The internal Microsoft email became public during the antitrust trial over the Oracle-PeopleSoft merger. Benjamin Pimentel, *E-mails Can Haunt Executives/ Unguarded Messages Can Show Up in Court*, SAN FRANCISCO CHRONICLE (July 5, 2004), <http://www.sfgate.com/cgi-bin/article.cgi?f=/c/a/2004/07/05/BUG267FNPV1.DTL>.

plementors—those who created software compatible with Windows—as competitors and either acquired or undermined them by integrating their product features into the Windows operating system itself. This practice discouraged developers and may have hampered evolution and innovation within the Windows ecosystem.

Like a biological ecosystem, business ecosystems undergo evolutionary processes of variation (new organizations, new knowledge), selection, and development. Firms following well-adapted strategies can survive and prosper. However, this biological analogy is not perfect. Unlike business systems, biological systems have no conscious intent and therefore evolution in biological systems is destiny. In business systems, this is not the case because economic agents, e.g., managers, entrepreneurs, investors, make conscious decisions and may have the opportunity to adjust a floundering strategy. So while there is path dependence, business ecosystems will reflect “evolution with design.”²² In other words, business and corporate “strategy processes are evolutionary by nature, but they involve significant elements of intentional design and orchestration of assets by managers.”²³

Business ecosystems are generally not exclusive in nature; firms may participate in more than one system. For example, personal computers using the Windows and Macintosh operating systems form the basis of two competing ecosystems, but Hewlett-Packard makes printers for both Windows and Macintosh users.

Within an ecosystem, the health and vitality of each firm is dependent on the health and vitality of all firms in the ecosystem, although some firms matter more than others. Some classic examples of where this condition applies are the shopping mall and the beehive. In each case, the demise of a key agent (the anchor tenant in a mall or the queen bee in a hive) can lead to the collapse of the whole system, even though the agent was not explicitly in charge of the system.

Whereas biological ecosystems are self-organizing, business ecosystems need not be, and frequently benefit from having an ecosystem manager, or “captain.” The ecosystem captain is a company that provides coordinating mechanisms, rules, key products, intellectual property, and financial capital to create structure and momentum for the market it seeks to create. When the captain is also a “platform leader,” the captain takes responsibility for guiding the technological evolution of the system to maintain competitiveness against rival ecosystems.²⁴

²² Mie Augier & David J. Teece, *Strategy as Evolution with Design: The Foundations of Dynamic Capabilities and the Role of Managers in the Economic System*, 29 *ORG. STUD.* 1187, 1188 (2008).

²³ *Id.* at 1201.

²⁴ ANNABELLE GAWER & MICHAEL A. CUSUMANO, *PLATFORM LEADERSHIP: HOW INTEL, MICROSOFT, AND CISCO DRIVE INDUSTRY INNOVATION* 245 (2002).

The role and identity of the captain within an ecosystem is not necessarily clear, and there may be multiple captains. In Japan's mobile telephony market, service operators such as NTT DoCoMo and KDDI are the captains who drive the ecosystem forward by deciding which handset makers to work with. In the United States, handset manufacturers, e.g., Apple, and content providers, e.g., Google, have recently emerged as captains who are able to affect the fortunes of U.S. service operators and other economic agents by their decisions of which ones they will work with.

An ecosystem requires rules for admission, entry, or both. Absent such rules, delicate complementarities can be disturbed and opportunities forsaken. With the involvement of numerous organizations, there are simply too many potential conflicts to allow for a completely self-organizing approach. Put differently, ecosystems are rife with externalities. However, individual companies should be free to not enter particular ecosystems. Tie-ins, agreements to bundle certain elements of a system together, are okay but tie-outs, agreements not to deal with certain other firms, may not be.²⁵ Closed, or semi-closed, ecosystems—sometimes called “walled gardens”²⁶—may promote innovation and are almost always socially desirable. Generally, market forces will oblige these walled gardens to compete with other ecosystems.

Apple's iPhone is an example of a semi-closed business ecosystem. Participation in the iPhone ecosystem requires recognizing Apple's intellectual property and abiding by Apple's rules. The Apple App Store, for example, requires application developers to grant Apple editorial control, including the right to disapprove of content. These rules are designed both to secure a superior customer experience and to protect Apple's business model. Apple's ongoing success in wireless products demonstrates that a tightly managed ecosystem can be as good as, or even superior to more flexible, “open” alternatives. Consumers are the ultimate beneficiaries of walled gardens. This is certainly true when they have choices; it may also be true when they do not.

From a public policy perspective, ecosystem-to-ecosystem rivalry—each featuring some degree of intra-ecosystem cooperation—should be recognized as a meaningful unit of analysis. As the Windows/Mac example demonstrates, ecosystems can compete fiercely with one another. The collective coordination of business strategies within an ecosystem sharpens and intensifies horizontal rivalry. In this sense, cooperation is the hand-

²⁵ Northern Pac. Ry. v. United States, 356 U.S. 1, 5-6 (1958).

²⁶ Thomas W. Hazlett et al., *Walled Garden Rivalry: The Creation of Mobile Network Ecosystems* 8 (Geo. Mason U. L. & Econ. Res. Paper Series, Working Paper No. 11-50 2011).

maiden of competition.²⁷ This has been recognized historically, as vertical relationships—integration—were seen to support horizontal competition.²⁸

Economic theory has yet to come to grips with ecosystems as a source of innovation. Increasingly, new markets emerge through a process of co-creation by a group of complementors, typically under the guidance of an ecosystem manager.²⁹ In economic theory, markets simply exist. The standard frameworks offer no scope for explaining the conscious, collective effort required to create markets and achieve a great deal of the innovation that characterizes the economy today.

IV. MODULARITY AND INTEGRATION

Modularity, which exists when the elements of a system interact with each other through “standardized interfaces within a standardized architecture,” is one of the underpinnings of ecosystems.³⁰ This section will focus on a comparison of modularity and vertical integration.

Standards-based modularity minimizes, if not eliminates, the need to transfer design and other information across organizational boundaries. These and other benefits of modularization, such as assisting in managing complexity and allowing innovation of complementary goods and services to proceed independently, are increasingly well understood.

The existence of a standard interface facilitates entry by complementors. It also helps to ensure that complements will be competitively supplied. This may not, however, always be in the system innovator’s best interest since it also enables imitators to replicate the product or system architecture. Ethiraj, Levinthal, and Roy suggest that a “near-modular” architecture, in which some interdependencies remain between modules, offers an optimal trade-off between ease of innovation and ease of imitation.³¹

Modularity theory originated in the 1960s, with the design theories of Herbert Simon.³² Interest in modularity was rekindled by the growth of the phenomenon and by Henderson and Clark, with their work on product ar-

²⁷ David J. Teece, *Competition, Cooperation, and Innovation: Organizational Arrangements for Regimes of Rapid Technological Progress*, 18 J. ECON. BEHAV. & ORG. 1, 12-13 (1992).

²⁸ Hazlett, *supra* note 26. See generally Frank H. Easterbrook, *The Limits of Antitrust*, 63 TEX. L. REV. 1 (1984) (explaining the Court history of vertical and horizontal relationships).

²⁹ Simone Scholten & Ulrich Scholten, *Platform-based Innovation Management: Directing External Innovational Efforts in Platform Ecosystems*, 3 J. KNOWLEDGE ECON. 164, 169-70 (2012).

³⁰ Richard N. Langlois, *Modularity in Technology and Organization*, 49 J. ECON. BEHAV. & ORG. 19, 19 (2002).

³¹ Sendil K. Ethiraj et al., *The Dual Role of Modularity: Innovation and Imitation*, 54 MGMT. SCI. 939, 940 (2008).

³² See Herbert A. Simon, *The Architecture of Complexity*, 106 PROC. AM. PHIL. SOC’Y 467, 467 (1962).

chitecture, and by Langlois and Robertson, with their work on (certain) innovative characteristics of industries based on compatibility standards and modular designs.³³ The modularity school is in part a challenge to Chandler's 1977 thesis that managerial hierarchies were necessary to coordinate large-scale productive systems.³⁴ Langlois went so far as to argue that, in the late 20th century, modular products and process architecture made hierarchical coordination unnecessary, leading him to re-characterize the managerial coordination that Chandler dubbed the "visible hand" as the "vanishing hand."³⁵

One should note, however, that networks which may appear modular from a distance are in fact more of a hybrid and are better thought of as "relational."³⁶ There are many examples of this type, particularly in the early stages of a market's emergence before modular standards are fully developed. However, this may only become evident upon close study of the network. Microsoft, for example, would seem to be the classic example of a modular provider of operating systems to the computer industry. But in the case of the new slim factor "ultrabooks," Microsoft has chosen to work closely with the computer manufacturers in order to improve the chassis design for these portable touchscreen devices in a way that makes the touchscreen interface software more reliable.³⁷ Microsoft does not operate at arm's length, as pure modularity would have it.

Classic vertical integration, in which all stages of production and distribution are coordinated within a single organization, has been in retreat as modular production networks have become dominant in a number of industries.³⁸ As Langlois has observed, since the late twentieth century, large, vertical firms have become "an increasingly small part of a landscape that features a wide variety of market and network forms."³⁹ Yet Langlois also acknowledges that vertically integrated enterprises will still be created "when circumstances dictate."⁴⁰ His thesis correctly focuses on the con-

³³ See Rebecca M. Henderson & Kim B. Clark, *Architectural Innovation: The Reconfiguration of Existing Product Technologies and the Failure of Established Firms*, ADMIN. SCI. Q., Mar. 1990, at 9; see also Richard N. Langlois & Paul L. Robertson, *Networks and Innovation in a Modular System: Lessons from the Microcomputer and Stereo Component Industries*, 21 RES. POL'Y 297, 297 (1992).

³⁴ CHANDLER, THE VISIBLE HAND, *supra* note 14, at 11.

³⁵ Richard N. Langlois, *The Vanishing Hand: The Changing Dynamics of Industrial Capitalism*, 12 INDUS. & CORP. CHANGE 351, 352 (2003).

³⁶ Gary Gereffi et al., *The Governance of Global Value Chains*, 12 REV. INT'L POL. ECON. 78, 83-84 (2005).

³⁷ Siu Han & Steve Shen, *Microsoft Stepping Up Involvement In Ultrabook Outer Designs*, Say Sources, DIGITIMES.COM (Apr. 16, 2012), <http://www.digitimes.com/news/a20120416PD213.html>.

³⁸ Timothy J. Sturgeon, *Modular Production Networks: A New American Model of Industrial Organization*, 11 INDUS. & CORP. CHANGE 451, 451-96 (2002).

³⁹ Langlois, *supra* note 35, at 353.

⁴⁰ *Id.*

junction of high throughput and thin markets as a driver of vertical integration.

In the early stages of an industry's evolution, certain inputs may not be available in competitive supply; in other words, markets are "thin." Vertical integration is then necessary to assure the quality or quantity of supply.⁴¹ As supplier capabilities, the number of suppliers, or both, increase, the advantages of vertical integration decline.⁴² This need not be a story about supplier capabilities alone—it can also be about supplier intent or willingness. When Qualcomm, a leading supplier of mobile phone technology, was founded in 1985, its CDMA technology was untested for cellular telephony, which led some of the leading telecom equipment suppliers to doubt it could work.⁴³ Faced with the prospect of limited support by suppliers, Qualcomm's management decided that it must offer an end-to-end solution of its own. In 1995 and 1996, when CDMA technology was first deployed in working networks, Qualcomm entered into the design and manufacture of infrastructure equipment, handsets, and the key chips that they require. At the same time, Qualcomm also licensed its technology to a wide variety of companies in order to improve the technology's supply base. In 1999, as CDMA began to gain traction in the market, Qualcomm exited the infrastructure and handset businesses.⁴⁴

Another important situation that will prevent the hand of management from vanishing anytime in the near future is that of systemic innovation. Systemic, or "architectural" innovation as it is sometimes called, requires coordinated development among a group of products composing a unified system.⁴⁵ Modularity theory holds that the ability of each element of the modularized system to advance at its own pace and to face direct competition leads to faster innovation than would a comparable integrated structure. This may hold when the technologies in the modules are truly autonomous. However, when a product requires, for superior performance, tight architectural integration of its elements, it may be more efficient—from both economic and innovation perspectives—to keep the design of all the elements

⁴¹ A similar argument is advanced in Richard N. Langlois, *The Capabilities of Industrial Capitalism*, 5 *CRITICAL REV.* 513, 513-30 (1991).

⁴² For a recent example of this type of industry evolution drawn from the electronics industry, see Timothy J. Sturgeon & Ji-Ren Lee, *Industry Co-Evolution: Electronics Contract Manufacturing in North American and Taiwan*, in *GLOBAL TAIWAN: BUILDING COMPETITIVE STRENGTHS IN A NEW INTERNATIONAL ECONOMY* (Suzanne Berger & Richard K. Lester eds., 2005).

⁴³ See Greg Linden, Clair Brown & Melissa M. Appleyard, *The Net World Order's Influence on Global Leadership in the Semiconductor Industry*, in *LOCATING GLOBAL ADVANTAGE: INDUSTRY DYNAMICS IN THE INTERNATIONAL ECONOMY* 244 (Martin Kenney & Richard Florida eds., 2004).

⁴⁴ Qualcomm stayed in the chip business rather than becoming a pure-play licensing firm. It has been one of the world's ten largest chip brands since 2008. It is one of the world's most innovative companies.

⁴⁵ Henderson & Clark *supra*, note 33, at 9-30; David J. Teece, *Economic Analysis and Strategic Management*, *CALIF. MGT. REV.*, Spring 1984, at 87, 102-04.

within a single organization. However, this need not extend to manufacturing; Apple designs the hardware and (operating) software for its highly integrated devices but does none of its own manufacturing.

As Apple's example suggests, vertical integration is not limited to the huge industrial enterprises of the previous century. Helper and Sako noted that, "For Chandler, the essence of 'the visible hand' was coordination within the managerial hierarchy, not vertical integration per se."⁴⁶ For example, Dell, a leading personal computer firm, tightly coordinates the activities of its suppliers by sharing order information and production plans in real time, leveraging the knowledge assets derived from its proprietary systems for balancing demand and supply. Dell also exerts influence by virtue of being a large customer, which helps it to create "virtual" integration with its network of independent suppliers.

Economic theory has not truly come to grips with vertical integration. For the most part, the mainstream theory implicitly assumes modularity. Economists assume that unbundling is not only viable but usually desirable. Marengo and Dosi have been critical—and rightfully so—of economists' willingness to ignore task interdependencies.⁴⁷ Furthermore, I too have made a similar criticism in articles opposed to the logic of telecom unbundling.⁴⁸

Transaction cost economics, as developed by Oliver Williamson, offers one explanation of vertical integration that has become widely accepted. In Williamson's framework, other things equal, when making outsourcing decisions, firms balance internal governance costs with (asset specificity-driven) transactions costs. However, when other things are (often) not equal, appropriability issues are likely to be paramount, and internal production costs and other manifestations of capabilities—including good corporate governance—may depend endogenously on the choice of market-based or internal organization.

Regulators have generally embraced the advantages of modularity without fully understanding its limitations. This has been most evident in the area of telecommunications networks, where unbundling mandates have risked harming innovation. By contrast, the courts have stopped short of imposing modularity through divestiture in the case against Microsoft. In considering whether an existing organizational architecture is anti-competitive, regulators must not only explore whether an economically

⁴⁶ Susan Helper & Mari Sako, *Management Innovation in Supply Chain: Appreciating Chandler in the Twenty-First Century*, 19 *INDUS. & CORP. CHANGE* 399, 415 (2010).

⁴⁷ Luigi Marengo & Giovanni Dosi, *Division of Labor, Organizational Coordination and Market Mechanisms in Collective Problem-Solving*, 58 *J. ECON. BEHAVIOR & ORG.* 303, 303-26 (2005).

⁴⁸ Thomas M. Jorde, J. Gregory Sidak & David J. Teece, *Innovation, Investment, and Unbundling*, 17 *YALE J. ON REG.* 1, 1-37 (2000); David J. Teece, *Telecommunications in Transition: Unbundling, Reintegration, and Competition*, 1 *MICH. TELECOMM. & TECH. L. REV.* 47, 47-78 (1995), available at <http://www.mttl.org/volone/teece.pdf>.

inefficient outcome has actually occurred, but whether a proposed intervention will compound inefficiencies.⁴⁹

V. CAPABILITIES AND ENTERPRISE GOVERNANCE

Economic theory and regulatory frameworks are also behind the times in understanding that the essence of the business enterprise lies in its capabilities. Economic models most often treat firms as homogeneous, leading the firms to make boundary decisions with reference only to transaction costs. Agency considerations animate choices with respect to financial structure.

A richer view of business organization is contained in the work of organization theorists and strategic management scholars. It is not that the mainstream views are wrong; they just capture too small a portion of the phenomenon at hand. The capabilities perspective is now well represented in most business schools, but the economics profession has not yet been impacted.

Ordinary capabilities, also known as competences, permit sufficiency, and sometimes excellence, in the performance of a delineated task. A firm's ordinary capabilities enable the production and sale of a defined, but static, set of products and services. Nevertheless, the presence of ordinary capabilities says nothing about whether the current production schedule is the right (or even a profitable) plan to follow. The nature of competences and their underlying processes, is such that they are not meant to change—until they have to.

The change process is a key element of higher-level competences called dynamic capabilities. Dynamic capabilities determine whether the enterprise is currently making the right products and addressing the right market segment(s). Dynamic capabilities are also forward-looking, helping to decide whether the enterprise's future plans are aligned with changing consumer needs and with technological and competitive opportunities.⁵⁰

Strong dynamic capabilities reflect an enterprise's excellence at orchestrating its resources, competences, and other assets. They allow the organization, especially its top management, to develop conjectures about the evolution of markets and technology, validate them, and realign assets and competences to meet new requirements. Dynamic capabilities are also used to assess when and how the enterprise is to ally with other enterprises and to engage in the co-creation of business ecosystems.

⁴⁹ Joseph Farrell & Philip J. Weiser, *Modularity, Vertical Integration, and Open Access Policies: Towards a Convergence of Antitrust and Regulation in the Internet Age*, 17 HARV. J.L. & TECH. 85, 85-134 (2003).

⁵⁰ David J. Teece et al., *Dynamic Capabilities and Strategic Management*, 18 STRATEGIC MGMT. J. 509, 515 (1997).

In short, the essential nature of dynamic capabilities is the astute sensing and seizing of opportunities, and then achieving the subsequent transformation of the enterprise as competitors crowd into the market. Top management plays a large role in these activities. Supporting routines and values must be deeply ingrained in the organization.

One place where policymakers have run afoul of the imperatives of enterprise capabilities is in the design of corporate governance mechanisms, specifically the composition of the board of directors. In the capabilities perspective, what matters most is the board's role in verifying that top management is pursuing a coherent strategic vision. In addition to the standard financial monitoring function, the board should also be responsible for responding to evidence of strategic malfeasance by management—cases where top management is making poor decisions with respect to the firm's changing environment.

Recent regulatory changes, such as the U.S. Sarbanes Oxley Act of 2002 (Sarbanes Oxley),⁵¹ have created greater financial transparency and require extremely tight financial controls and rigorous—some might say pedantic—application of accounting rules. However, this type of rigor and oversight provides little protection against strategic blunders by management. Indeed, by focusing board attention elsewhere, Sarbanes Oxley is likely to amplify the likelihood of such blunders.

The new technical requirements of good governance now prioritized in U.S. law may be of only second or third-order importance relative to the larger issues that truly good governance requires. Furthermore, what constitutes “good governance” may, in fact, be context-dependent. For example, in some circumstances, the separation of the roles of CEO and chairman may be counter-productive to the rapid transformation required to meet a competitive threat, or to develop and commercialize a new technology that is meeting resistance from certain parts of the company. Bifurcated responsibilities and decision rights might well complicate leadership issues and slow transformation.

Many corporate boards today may have insufficient strength to help management properly evaluate strategic alternatives. Board members typically lack staff to conduct their own analyses, leaving them reliant on themselves and management for their understanding of complex issues. In the contemporary governance environment, greater weight has been placed on the need for board members who are independent of management, but not on members who understand the industry environment in which the company must compete.⁵²

⁵¹ See Sarbanes-Oxley Act of 2002, Pub. L. No. 107-204, 116 Stat. 745 (codified as amended in scattered sections of 11, 15, 18, 28 and 29 U.S.C.).

⁵² David J. Teece, *Management and Governance of the Business Enterprise: Agency, Contracting, and Capabilities Perspectives*, in THE OXFORD HANDBOOK OF CAPITALISM (Dennis C. Mueller ed., 2012).

In short, rigid rules such as those imposed by the stock exchanges in the U.S. in reaction to the Enron debacle and other scandals arguably hurt more than they help in the provision of good governance. In fact, the New York Stock Exchange's mandate that boards have more independent members than management members—which goes beyond the Sarbanes Oxley independence requirement that applies only to the audit committee—may have weakened governance in the areas where it matters most.

The antitrust implications of the capabilities framework may be equally far-reaching. The capabilities approach implicitly explains firm-level heterogeneity, but concentration indices and other measures of market power implicitly assume homogeneity. They focus on product markets when the real action that defines competition lies upstream, embedded in the firms' capabilities.

In conventional analysis, market share is viewed as an inverse measure of competitive vulnerability. There is no reason, however, that this need be the case, since a large market share can derive from chance, network effects, or other factors that are not dependent on the firm's innate capabilities. Andy Grove, Intel's former CEO, once claimed, quite correctly, that all Intel's relatively high market share provides is a seat at the table for the next round of innovation. Incumbency conveys no special benefits in regimes of rapid technological change. In fact, large incumbents are often more vulnerable because they are often reluctant to implement innovations that would compete with their existing products—innovations that are readily adopted by rivals and new entrants.⁵³ Therefore, there is no substitute for assessing competition at the capabilities level. Analyzing product market shares and treating them as a measure of competitive vulnerability are often quite meaningless. Yet this is the very essence of how competition authorities around the world implement competition policy.

VI. THE MULTI-INVENTION CONTEXT

Yet another place where mainstream economic theory has fallen short of developments in the business world is intellectual property. Textbook treatments of the innovation—invention nexus often assume that products may “read on” one or a few patents, not thousands. In fact, it is often the case, and has been true for years, that complicated products—particularly those with many components, parts or functions—may read on hundreds, if not thousands, of patents. The smartphone is a contemporary example of this “multi-invention context.”⁵⁴

⁵³ CLAYTON M. CHRISTENSEN, *THE INNOVATOR'S DILEMMA* 108 (1997).

⁵⁴ Deepak Somaya et. al., *Innovation in Multi-Invention Contexts: Mapping Solutions to Technological and Intellectual Property Complexity*, 53 CAL. MGMT. REV. 47, 48-50 (2007).

The sewing machine provides what is perhaps the earliest example of a case in which the ownership of patents by a number of different entities required complex licensing. The first U.S. patent for a sewing machine was issued in 1842, and the first Singer Sewing Machine went on sale in 1850. Adam Mossoff has shown that, by 1855, there were seventy U.S. patents on sewing machines, and the patent holders engaged in years of high-profile litigation and attempts to claim misattribution of patents at the U.S. Patent Office, before forming a patent pool in 1856.⁵⁵

Patent pools are one, but by no means the only, way of managing the multi-invention challenge, at least in the limited cases where they can be formed. Cross-licensing is usually the more important mechanism for achieving patent peace.⁵⁶ In a patent pool, two or more companies combine their patent rights and offer one-stop licensing, often at set royalty rates. This has the advantage for the licensees of lowering the transaction costs of entering the industry. For the licensors, it provides an expanded opportunity to monetize their intellectual property. Recent examples of patent pools include the RFID Consortium, for radio frequency identification (RFID) technology in the now-ubiquitous electronic tags and readers used for inventory management, and MPEG LA LLC, for the audio and video compression standards that permit multimedia content to be transmitted efficiently over the Internet.

The creation of the Sewing Machine Combination in 1856 unleashed an explosion of innovation. By 1862, over 350 patents had been granted on sewing machine improvements and accessories.⁵⁷ The market for sewing machines grew astonishingly fast, despite the litigation and a cultural bias against women's use of mechanical devices. Similarly, innovation in mobile phones continues to be remarkably robust. Hard-fought patent disputes involving many of the major industry participants have not slowed innovation.

The mobile phone industry is an important contemporary example of a multi-invention environment. Mobile handsets only work if they precisely match the functional and communication protocols displayed by the network carrier. These products evolve over time, but not instantly; standards stay frozen for at least certain periods. Any product compatible with a standard—e.g., GSM, UMTS, LTE, or the CDMA family—necessarily implements the standard's technical specifications, which detail the protocol, data format signal, and other matters required for a handset to communicate with a base station to complete a call. These standards are established by a standard setting organization (SSO) and are comprised of indus-

⁵⁵ Adam Mossoff, *The Rise and Fall of the First America Patent Thicket: The Sewing Machine Wave of the 1850s*, 53 ARIZ. L. REV. 165, 166 (2011).

⁵⁶ Peter C. Grindley & David J. Teece, *Managing Intellectual Capital: Licensing and Cross-Licensing in Electronics*, 39 CAL. MGMT. REVIEW 8, 18 (1997).

⁵⁷ Mossoff, *supra* note 55, at 194.

try participants' technical solutions embedded in standards covered by patents. Patents embedded within a standard are referred to as essential, if they are indeed essential to practicing the standard. However, there is no arbitrator within the SSO of what is "essential;" it is a self-declared adjective.

Most standard-setting bodies require owners of essential patents to make licenses available on fair, reasonable, and non-discriminatory (FRAND) terms. The parties must negotiate the specific terms of such licenses. Cumulative royalty rates for complicated products that integrate multiple technologies, often on a single silicon chip, can be quite high if the infringer does not have a patent portfolio to cross-license. With a cross-license, the balancing payments may be small, possibly zero.

A new entrant must endeavor to invent and patent as quickly and robustly as possible in order to improve the terms it will face for the cross-licensing agreements needed to secure design and operating freedom. Balancing payments of cash from those with less valuable patents to those with a more valuable portfolio ensures that "free riding" does not occur. The payments also make new entry possible—a possibility that would be denied absent FRAND commitments. Balancing payments also ensure that firms contributing a disproportionate value to the industry's stock of patented inventions earn a reasonable return on their investments and efforts.

Some economists and policymakers respond to the growing prevalence of multi-invention situations by calling for weaker patent rights, as this lowers the cost for new entrants that have contributed less (or not at all) to the industry's knowledge base. This is likely to be bad public policy. It reduces the incentives for all innovators and penalizes those firms and individuals that have contributed to the invention pool. It favors the here and now, over the future. It penalizes innovating nations and favors imitators.

More than a century of experience has shown—e.g., with sewing machines—that private ordering—i.e., privately negotiated—arrangements work. Innovators in multi-invention contexts can cope with the requirement to scan for, negotiate over, and license in other innovations as needed. Multiple generations of wireless technology have been launched quickly without disruption. Licensing costs have not prevented rapid innovation in electronics, biotechnology, medical products, and other domains.

Policymakers must be especially careful not to create additional uncertainties through clumsy policy interventions. Uncertainty about patent validity, scope, and applicability will undermine private ordering and take transactions out of the marketplace and into the courthouse, with negative consequences likely for all but the lawyers. Eliminating the right to obtain injunctions will have similar effects.⁵⁸

⁵⁸ Brief for Respondent at 2, *eBay, Inc. v. MercExchange, L.L.C.*, 547 U.S. 388 (2006) (No. 05-130). David Teece is an additional co-author of the document but is not listed as counsel on it because he is not a lawyer.

The United States in the early 19th century was, in fact, a pioneer in recognizing that some amount of exclusivity stemming from patent rights spurs invention and innovation. This recognition provided momentum to the Second U.S. Industrial Revolution.⁵⁹ However, since the Sherman Antitrust Act of 1890, there is a serious risk that the law will be used to limit the behavior of successful innovators with antitrust consents that require compulsory licensing or even the royalty-free transfer of know-how. For example, the Federal Trade Commission's 1975 settlement with Xerox forced the company not only to license its patents to rivals at predetermined rates but also to provide them access to its written know-how. This effectively "subsidized" the entry into the U.S. market of Japanese competitors just as they were preparing to make inroads in the plain-paper copier market.⁶⁰

It is not uncommon for infringers to try to avoid paying licensing fees for the use of others' technologies by filing antitrust cases. Judges and juries would do well to deny such claims and to avoid undermining intellectual property rights. To do otherwise risks undermining future innovation and growth. Antitrust intervention applied injudiciously increases uncertainty and harms consumers, especially the next-generation consumers.

CONCLUSION

In this essay, as well as in the other presentations at the Digital Inventor Conference, a number of aspects of competition and innovation have been reviewed. Many of them have been around in some form for a long time: multi-invention contexts, firm-level capabilities, and modularity. However, taken together, these aspects of the current global business landscape, combined with advances in information technology and communications, place businesses in a radically altered environment from just twenty years ago.

The digital realm will keep creating new business models that challenge the established tenets of regulatory policy. Multi-sided platforms are a case in point. The rules of thumb for assessing proper conduct, such as marginal cost pricing benchmarks, do not match the economic reality of operating a multi-sided platform.⁶¹

Scholars are now scrambling—or should be scrambling—to incorporate next-generation competition into their frameworks. Literature on the multi-patent situation, cross-licensing, patent pools, multi-sided platforms,

⁵⁹ B. Zorina Khan, *Antitrust And Innovation Before The Sherman Act*, 77 ANTITRUST L.J. 757, 758-59 (2011).

⁶⁰ F. M. Scherer, *The Role of Patents in two US Monopolization Cases*, 12 INT'L J. BUS. ECON 297, 304 (2005).

⁶¹ David S. Evans, *The Antitrust Economics of Multi-Sided Platform Markets*, 20 YALE J. ON REG. 325 (2003).

and modularity is increasing rapidly. However, research on business ecosystems lags. There is a vibrant literature on capabilities in the strategic management field, but it has yet to sufficiently impact economic research, competition policy, corporate governance, and public policy more generally. The law is also slow to grasp the impact of these recent developments.

Policymakers and regulators must encompass next-generation competition in their analyses or risk deeper policy error. However, they can take comfort in the fact that next-generation competition is relentless, provided the innovation engine keeps firing. The best economic policies and legal structures will almost always be those that respect intellectual property rights, promote innovation, allow private ordering arrangements, and favor the future.

COLLATERAL CONSEQUENCES OF CRIMINAL CONVICTIONS:
A COST–BENEFIT ANALYSIS

*Genevieve J. Miller**

INTRODUCTION

“Sam, age 22,” is drinking at a party and meets a girl, “Lisa, age 21.”¹ Lisa is drinking as well and becomes very intoxicated. Sam and Lisa flirt for a while, and Sam thinks she likes him. Lisa tells Sam she is going upstairs and suggests he join her. Sam goes up a few minutes later and finds her lying down on the bed. Sam begins performing oral sex on her when Lisa’s friend walks in and subsequently calls the police. Sam is charged with sexual assault because of Lisa’s incapacitated state. He negotiates a plea agreement that results in a felony conviction for aggravated sexual battery.² Since Sam has no prior criminal history, he serves a short jail sentence and then completes three years of probation, which includes sex offender treatment. Fifteen years later, Sam is married, the proud father of two young children, and gainfully employed. Sam has had no other contact with the court system save for a non-reckless speeding ticket he received five years ago. He wants to attend a school assembly to see his oldest son receive an award for his science project—an assembly to which all parents with children receiving awards have been invited. It will be held at 6:00 PM in the school’s auditorium. Unfortunately, Sam has to tell his son he cannot go because of his conviction fifteen years ago, even though that incident did not involve a minor. Sam is barred from going onto the premises of any school for a school function—even if the function is for his own child, geared towards parents, and after regular school hours.³

Consequences stemming from a criminal conviction can be found in less vivid circumstances as well. For example, a woman who cleans houses

* J.D. Candidate, 2013, George Mason University School of Law; B.A. with distinction, University of Virginia. The author extends her gratitude to Bonnie Hoffman, Loudoun, Fauquier, and Rappahannock Counties Deputy Public Defender, for her invaluable insight and guidance as a mentor, and to the author’s friends and family for their constructive edits of this comment.

¹ The facts in the following scenario are modified from a case handled by the Loudoun County, Virginia, Office of the Public Defender. Names and some details have been changed.

² *E.g.*, VA. CODE ANN. § 18.2-67.3 (West 2011).

³ Many jurisdictions ban those convicted of certain sex offenses from entering school premises at specific times. *See, e.g.*, ARK. CODE ANN. § 5-14-128 (West 2011); GA. CODE ANN. § 42-1-15 (West 2011); 720 ILL. COMP. STAT. ANN. 5/11-9.3 (West 2011); IND. CODE ANN. § 35-38-2-2.2 (West 2008); IOWA CODE ANN. § 692A.113 (West 2011); OKLA. STAT. ANN. tit. 57, § 590 (West 2011); VA. CODE ANN. § 18.2-370.5 (West 2011).

for a living is injured at work.⁴ She receives a prescription for pain medication and, growing increasingly tolerant of the medication, becomes addicted to it. In order to deal with the addiction and tolerance, she begins altering her doctor's prescription, adding refills that were not authorized and changing the number of pills per prescription. She is caught and faces criminal charges for prescription fraud.⁵ Accepting a plea bargain, she is convicted of one charge, rather than the multiple counts she is facing from the weekly prescriptions she altered and passed. The agreement allows her to avoid active jail or prison time, instead placing her on probation. While on probation, she completes a substance abuse treatment program. Five years later, she wants to go to college to become a social worker, but her prior conviction makes her ineligible for federal student loans.⁶ Without student loans, she is unable to afford school. Regardless of her successful rehabilitation, the long-lasting consequences of her plea create significant barriers to future employment through both the denial of access to loans and additional restrictions.⁷

Collateral consequences, also known as collateral sanctions or civil disabilities, are those penalties that attach to a criminal conviction, whether a misdemeanor or felony, even if the sanction is not included in the sentence.⁸ Frequently, those charged with criminal offenses do not know or fully appreciate the panoply of consequences they will face if they are convicted.⁹ Furthermore, as these consequences are not subject to constitutional ex post facto considerations, even the most industrious and astute attorneys and defendants cannot fully contemplate the potential future consequences of a conviction.¹⁰ "Collateral consequences can operate as a secret sentence."¹¹

Most of the time, defendants, their counsel, prosecutors, and courts only focus on the most obvious results of a conviction—the time a defendant

⁴ The facts in the following scenario are modified from a case handled by the Loudoun County, VA, Office of the Public Defender. Names have been omitted and some details have been changed.

⁵ *E.g.*, VA. CODE ANN. § 18.2-258.1 (West 2011).

⁶ 20 U.S.C. § 1091(r) (2006).

⁷ Leroy D. Clark, *A Civil Rights Task: Removing Barriers to Employment of Ex-Convicts*, 38 U.S.F. L. REV. 193, 195-97 (2004) (describing generally the types of jobs that are not open to ex-offenders, including professions requiring a license, public employment with the state and federal governments, and private employers, one-third of whom routinely run criminal background checks in order to weed out ex-offenders).

⁸ STANDARDS FOR CRIMINAL JUSTICE § 19-1.1 (2004), available at http://www.americanbar.org/publications/criminal_justice_section_archive/crimjust_standards_collateral_blk.html#1.1.

⁹ See Gabriel J. Chin & Richard W. Holmes, Jr., *Effective Assistance of Counsel and the Consequences of Guilty Pleas*, 87 CORNELL L. REV. 697, 700 (2002).

¹⁰ Because collateral consequences are not deemed a direct sanction resulting from a criminal conviction, they are not subject to the requirements of U.S. CONST. art. I, § 9, cl. 3 or U.S. CONST. art. I, § 10, cl. 1. See *Calder v. Bull*, 3 U.S. 386, 390 (1798).

¹¹ Chin & Holmes, Jr., *supra* note 9.

may spend in jail or in prison, the amercing of a fine, or the length and conditions of probation.¹² Often lost in these discussions are the long-lasting, far-reaching, and usually more life-altering collateral consequences that follow these criminal convictions.¹³ Significant consequences can occur regardless of whether an individual is convicted of a felony or a misdemeanor, or whether they serve years in the penitentiary or do not serve a single day in jail.¹⁴

Collateral consequences come in a myriad of forms. They include temporary or permanent ineligibility for social security or food stamp benefits because of a drug conviction,¹⁵ and loss of government-assisted housing because of either specific drug or alcohol convictions¹⁶ or general criminal activity.¹⁷ Other collateral sanctions can bar an ex-offender from obtaining certain professional licenses, employment with federal and state agencies, and employment in the private sector.¹⁸ Still other sanctions include disqualification from military enlistment,¹⁹ disenfranchisement,²⁰ and ineligibility for jury service.²¹ Individuals who have been convicted of a drug offense are also barred from receiving federal student educational assistance.²² Long-time immigrants lawfully in the United States face the very real pos-

¹² See ROBERT C. BORUCHOWITZ, MALIA N. BRINK & MAUREEN DIMINO, NAT'L ASS'N OF CRIMINAL DEF. LAWYERS, *MINOR CRIMES, MASSIVE WASTE: THE TERRIBLE TOLL OF AMERICA'S BROKEN MISDEMEANOR COURTS* 34 (2009).

¹³ *Id.*

¹⁴ Michael Pinard, *An Integrated Perspective on the Collateral Consequences of Criminal Convictions and Reentry Issues Faced by Formerly Incarcerated Individuals*, 86 B.U. L. REV. 623, 635 (2006).

¹⁵ 21 U.S.C. § 862a (2006); see e.g. U.S. DEP'T OF JUSTICE, OFFICE OF JUSTICE PROGRAMS, BUREAU OF JUSTICE ASSISTANCE, *DENIAL OF FEDERAL BENEFITS PROGRAM AND CLEARINGHOUSE* 3 (2002), available at <https://www.ncjrs.gov/pdffiles1/bja/193770.pdf> (noting that social security benefits can be denied for up to five years following an individual's first drug distribution offense, and up to one year following a drug possession offense; this denial of benefits is not based on how much the individual has paid into the system).

¹⁶ 42 U.S.C. § 13661 (2006).

¹⁷ *Id.* § 1437f(d)(1)(B)(iii).

¹⁸ Clark, *supra* note 7 (describing generally the types of jobs that are not open to ex-offenders).

¹⁹ 10 U.S.C. § 504(a) (2006).

²⁰ Voting restrictions are based on state law. See THE SENTENCING PROJECT, *FELONY DISENFRANCHISEMENT LAWS IN THE UNITED STATES* 1-3 (2011), available at http://www.sentencingproject.org/doc/publications/publications/fd_bs_fdlawsinus_Sep2012.pdf.

²¹ Brian C. Kalt, *The Exclusion of Felons from Jury Service*, 53 AM. U. L. REV. 65, 67, 150-57 (2003) (observing that the federal government and the following thirty-one states permanently exclude convicted felons from jury service: Alabama, Arkansas, California, Delaware, Florida, Georgia, Hawaii, Kentucky, Louisiana, Maryland, Michigan, Mississippi, Missouri, Montana, Nebraska, Nevada, New Hampshire, New Jersey, New Mexico, New York, Ohio, Oklahoma, Pennsylvania, South Carolina, Tennessee, Texas, Utah, Vermont, Virginia, West Virginia, and Wyoming).

²² 20 U.S.C. § 1091(r) (2006).

sibility of deportation²³ and bars to reentry²⁴ as a collateral consequence of a conviction.

This comment explores the variety of collateral consequences faced by those convicted of a crime. Part I traces the development and increased impact of collateral consequences, focusing on the pressures that attorneys and clients face in navigating the legal system. Part II examines the economic ramifications of collateral consequences by exploring the costs and benefits of such consequences and the externalities these consequences create. The analysis presented demonstrates that the costs of collateral consequences often outweigh the benefits society derives from further sanctions against those convicted of crimes, and shows collateral consequences create significant negative externalities. Part III examines collateral sanctions in the case of female ex-offenders, a group particularly disadvantaged by the status quo. Finally, Part IV offers solutions to ameliorate the often devastating economic effects of collateral sanctions, specifically decriminalization, tying collateral consequences to ex post facto considerations, and legislation advocating expungement for first-time offenders.

II. HISTORY AND CURRENT IMPACT OF COLLATERAL CONSEQUENCES

A. *History of Collateral Consequences*

Formal collateral consequences resulting from state-sanctioned punishment can be traced back to ancient Greece and Rome.²⁵ In ancient Greece, civil disabilities prohibited a criminal pronounced “infamous” from appearing in court, voting, making speeches, attending assemblies, and serving in the army.²⁶ Romans later adopted these same civil disabilities.²⁷ By 1066, England had adopted an analogous system of civil disabilities in which an “attainted” criminal theoretically, but not always in practice, lost property rights and all civil rights.²⁸ These sanctions sought to further the goals of retribution and deterrence by imposing severe punishments for convicted criminals, thereby encouraging others to abide by the law.²⁹ The

²³ See generally 8 U.S.C. § 1227 (2006); Guy Cohen, Note, *Weakness of the Collateral Consequences Doctrine: Counsel's Duty to Inform Aliens of the Deportation Consequences of Guilty Pleas*, 16 FORDHAM INT'L L.J. 1094, 1111-12 (1993) (discussing convictions that expose aliens to deportation).

²⁴ See, e.g., 8 U.S.C. § 1182 (2006) (listing a number of different grounds for which an alien may be barred from entering the United States).

²⁵ Walter Matthews Grant et al., Special Project, *The Collateral Consequences of a Criminal Conviction*, 23 VAND. L. REV. 929, 941-42 (1970).

²⁶ *Id.* at 941.

²⁷ *Id.* at 942.

²⁸ *Id.* at 942-43 (explaining that in even the harsh sanctions present in medieval England, statutes were frequently not as severe in practice).

²⁹ *Id.* at 944.

American penal system, heavily influenced by English law, adopted a system of civil disabilities with similar aims.³⁰

Criminal law in the United States has always included civil disabilities to varying degrees.³¹ During the 1980s and 1990s, the variety and severity of collateral sanctions rapidly increased due to the proliferation of the “tough on crime” and “war on drugs” movements advanced by politicians.³² Partially because of this focus on crime, incarceration levels increased dramatically. In 1973, 200,000 people were incarcerated.³³ In 2003, that number rose to 1.4 million,³⁴ a 600% increase. In contrast, the United States population over that same period only grew by 38%.³⁵ Thus, incarceration rates increased 15.79 times faster than population growth.³⁶ Regardless of whether those convicted of a crime actually serve time in jail, the collateral consequences today’s generation of offenders face—sanctions which were passed into law with little to no focus on their hidden financial impact—are more punitive and less individualized than those of the past few decades.³⁷ The increase in incarcerations coupled with more severe penalties demonstrates that more people than ever are subject to the barriers imposed by collateral sanctions.

With the increase in the type and severity of collateral consequences, society has become more attuned to the repercussions of such sanctions. Compilations of federal statutes which impose sanctions for convicted felons are readily accessible online.³⁸ Academics have scrutinized the effect

³⁰ *Id.* at 949 (describing the goals of retribution and deterrence in both countries’ penal systems).

³¹ Grant et al., *supra* note 25, at 949-51.

³² Pinard, *supra* note 14, at 637. The Republican Party is largely credited with the imposition of tougher sanctions for crimes beginning in the 1960s and extending through the 1990s. Paul Parker, *A Review of The Politics of Injustice: Crime and Punishment in America*, 12 J. OF CRIM. JUST. & POPULAR CULTURE 71, 72-73 (2005) (book review). Some academics have observed that these movements were more successful as political tools rather than as deterrents to crime. *Id.* at 74.

³³ JEREMY TRAVIS, BUT THEY ALL COME BACK: FACING THE CHALLENGES OF PRISONER REENTRY 23 (2005).

³⁴ *Id.*

³⁵ *United States Population Growth*, CENSUSSCOPE.ORG, http://www.censusscope.org/us/chart_popl.html (last visited Jan. 7, 2012). This percentage was calculated by taking the total U.S. population in 2000 (281,421,906), subtracting the total U.S. population in 1970 (203,302,031), dividing the difference (78,119,875) by the total U.S. population in 1970, and multiplying by 100.

³⁶ This value was derived by dividing 600% by 38%.

³⁷ Pinard, *supra* note 14, at 637; ABA STANDARDS FOR CRIMINAL JUSTICE, COLLATERAL SANCTIONS AND DISCRETIONARY DISQUALIFICATION OF CONVICTED PERSONS 8 (3d ed. 2004), available at http://www.americanbar.org/content/dam/aba/publishing/criminal_justice_section_newsletter/crimjust_standards_collateralsanctionwithcommentary.authcheckdam.pdf; Nora V. Demleitner, *Preventing Internal Exile: The Need for Restrictions on Collateral Sentencing Consequences*, 11 STAN. L. & POL’Y REV. 153, 155 (1999).

³⁸ U.S. DEP’T OF JUSTICE, FEDERAL STATUTES IMPOSING COLLATERAL CONSEQUENCES UPON CONVICTION (2000), available at https://docs.google.com/viewer?a=v&q=cache:5oioaltcMqkJ:www.justice.gov/pardon/collateral_conseq

of some collateral consequences such as felon disenfranchisement.³⁹ Perhaps most significantly, the U.S. Supreme Court's recent decision in *Padilla v. Kentucky* heightened the legal profession's focus on collateral sanctions.⁴⁰ *Padilla's* holding solidified the prior trend of states requiring that judges and attorneys warn non-citizen defendants of the potential immigration consequences of criminal convictions.⁴¹ At least in the case of immigration concerns, attorneys are now required to explain to their clients this potential collateral consequence.⁴²

While *Padilla* has shed light on the immigration ramifications of criminal proceedings, some jurisdictions have declined to address, or have outright resisted, extending the notification requirement outside the context of immigration.⁴³ Greater focus on collateral consequences has caused tension between an attorney's obligations to fully inform a client of all the potential effects of a conviction, and the increasingly difficult task of staying abreast of the continuously changing and increasing collateral consequences of criminal convictions.⁴⁴ This tension is especially significant in light of overflowing criminal court dockets.

uences.pdf+&hl=en&gl=us&pid=bl&srcid=ADGEESjADONpILsnCiwR_P2m6eUEzakOwnQA-BvQfKcf44Zb8psJtNU_HD8aP3o5hP44Zni6EhdzmWIF8F4pcTDielodSh54tPvHtRGDX59stfwpj5TsUDciHUfKeXedtKm2J0TnrN&sig=AHIEtbTiUkJMZqBbh_KDPa0MYf_3rRE4jA.

³⁹ See generally Clark, *supra* note 7, at 201; Nora V. Demleitner, *Continuing Payment on One's Debt to Society: The German Model of Felon Disenfranchisement as an Alternative*, 84 MINN. L. REV. 753, 754-56 (2000); Matthew E. Feinberg, *Suffering Without Suffrage: Why Felon Disenfranchisement Constitutes Vote Denial Under Section Two of the Voting Rights Act*, 8 HASTINGS RACE & POVERTY L.J. 61, 61-65 (2011); Elena Saxonhouse, Note, *Unequal Protection: Comparing Former Felons' Challenges to Disenfranchisement and Employment Discrimination*, 56 STAN. L. REV. 1597, 1598-1601 (2004).

⁴⁰ *Padilla v. Kentucky*, 130 S. Ct. 1473, 1483 (2010).

⁴¹ See, e.g., Pinard, *supra* note 14, at 644.

⁴² *Padilla*, 130 S. Ct. at 1482.

⁴³ *People v. Hughes*, No. 2-09-0992, 2011 Ill. App. 2d 090992, at *6 (Ill. App. Ct. July 19, 2011) (declining to extend *Padilla* notification requirements for collateral consequences that resulted from a defendant's conviction as a sexually violent predator); *Thomas v. United States*, Civil Action No. RWT-10-2274, Criminal No. PMD-06-4572, 2011 WL 1457917, at *4 (D. Md. Apr. 15, 2011) (denying the defendant's argument that he had a right to be notified of the employment-related collateral consequences of conviction under the rationale set forth in *Padilla*); *State v. Rasheed*, 340 S.W.3d 280, 284 n.1 (Mo. Ct. App. 2011) (declining to address whether *Padilla's* holding should be extended outside of the deportation context); *Maxwell v. Larkins*, No. 4:08 CV 1896 DDN, 2010 WL 2680333, at *10 (E.D. Mo. July 1, 2010) (holding that *Padilla* did not require an attorney to warn the defendant of potential civil commitment under the Sexually Violent Predator Act).

⁴⁴ Renee Newman Knake, *The Supreme Court's Increased Attention to the Law of Lawyering: Mere Coincidence or Something More?*, 59 AM. U. L. REV. 1499, 1567-69 (2010).

B. *Overloaded Dockets and Coercion: Forcing Defendants to Plea*

Most criminal cases do not go to trial on their merits, as the vast majority of criminal defendants in federal and state courts plead guilty, usually in connection with some type of plea agreement with the prosecutor.⁴⁵ In 2004, roughly 86%—71,692 out of 83,391—of federal defendants pleaded guilty.⁴⁶ In that same year, only 4%—3,346 out of 83,391—of federal criminal cases went to trial.⁴⁷ Similarly, only 5% of all state felony criminal prosecutions went to trial in 2004.⁴⁸

The fact that so few cases actually go to trial does not diminish the vital role of effective assistance of counsel for most defendants. Research demonstrates that when faced with major decisions, individuals frequently put greater weight on the risk of an unfavorable outcome without due consideration for the probability or likelihood that the risk will occur.⁴⁹ For instance, people focus on emotional cues and stress when making high-stakes decisions with uncertain outcomes.⁵⁰ These biases that skew decision-making are indoctrinated early on, as studies show that even children and adolescents disregard the probabilities of various outcomes when making decisions under uncertain circumstances.⁵¹ Extrapolating from this fact, it follows that defendants, trying to determine whether to accept a plea or go to trial, will focus on the *severity* of a consequence, while failing to consider the *probability* of that consequence, and therefore are likely heavily influenced into accepting a plea.⁵² In light of the existence of such a bias—and given the constitutional guarantee of the right to counsel⁵³—it is that much more important to assure that defendants have access to adequate and

⁴⁵ See U.S. DEP'T OF JUSTICE, BUREAU OF JUSTICE STATISTICS, SOURCEBOOK OF CRIMINAL JUSTICE STATISTICS ONLINE Table 5.17.2004 (2004), available at <http://www.albany.edu/sourcebook/pdf/t5172004.pdf>; *id.* Table 5.46.2004, available at <http://www.albany.edu/sourcebook/pdf/t5462004.pdf>.

⁴⁶ See *id.* Table 5.17.2004.

⁴⁷ *Id.*

⁴⁸ *Id.* Table 5.46.2004.

⁴⁹ Howard Kunreuther et al., *High Stakes Decision Making: Normative, Descriptive and Prescriptive Considerations*, 13 MKTG. LETTERS 259, 261 (2002).

⁵⁰ *Id.* at 262-63.

⁵¹ Jonathan Baron et al., *Decision Making Biases in Children and Early Adolescents: Exploratory Studies*, 39 MERRILL PALMER Q. 23, 19-20 (1993), available at <http://www.sas.upenn.edu/~baron/papers/kdm.pdf> (pagination refers to the online document).

⁵² Kunreuther et al., *supra* note 49, at 261.

⁵³ See U.S. CONST. amend. VI (guaranteeing the accused the right to counsel in all criminal proceedings); *Gideon v. Wainwright*, 372 U.S. 335, 344 (1963) (requiring, in the interest of fairness, that states provide counsel to defendants who cannot afford representation given the severe consequences that can result from our adversarial system of justice); *Argersinger v. Hamlin*, 407 U.S. 25, 38 (1972) (extending the right to counsel to misdemeanor defendants).

effective representation at the earliest stages of a criminal case, as this is the time when many cases are often resolved.

Prior to *Padilla*, neither the judge nor defense counsel were required to inform a defendant of the existence of any collateral consequences.⁵⁴ It is not clear what impact, if any, *Padilla*'s holding will have on the frequency of plea bargains.⁵⁵ It may be the case that future rulings will extend *Padilla* beyond the deportation context and require that defendants be informed of the additional consequences of a guilty plea.⁵⁶ While professional standards and ethical obligations provide that an attorney should notify a defendant of collateral sanctions,⁵⁷ it remains to be seen whether courts will uniformly extend this right to all defendants.⁵⁸ Regardless of *Padilla*'s impact, the fact remains that the courts, and by extension attorneys, are overworked and therefore unable to both fully investigate the charges against a defendant and educate that defendant of all the potential collateral consequences of a conviction or plea.⁵⁹

The criminal justice system is becoming increasingly strained, as each year legislatures enact more laws criminalizing behavior and more individuals are prosecuted.⁶⁰ The impact of this strain is felt at every level, from prosecutors to defense attorneys, and from courts to jails.⁶¹ Misdemeanor courts suffer the greatest burden.⁶² Not only does increased criminalization mean that more defendants are being charged and prosecuted, thereby straining the system's limited resources, but additionally there is a belief that because these matters are *only* misdemeanors, they deserve less time and attention than felony cases.⁶³ Nevertheless, the vast majority of indi-

⁵⁴ Clark, *supra* note 7, at 197.

⁵⁵ Michael Vomacka, *Supreme Court Decisions in Padilla v. Kentucky States Affirmative Duty to Inform Client of Risk Guilty Plea May Result in Removal*, 25 GEO. IMMIGR. L.J. 233, 235 (2010).

⁵⁶ Gabriel J. Chin & Margaret Love, *Status as Punishment: A Critical Guide to Padilla v. Kentucky*, CRIM. JUST., Fall 2010, at 22.

⁵⁷ STANDARDS FOR CRIMINAL JUSTICE § 14-1.4 (1999), available at http://www.americanbar.org/publications/criminal_justice_section_archive/crimjust_standards_guiltyplea_as_blk.html#1.4.

⁵⁸ Chin & Love, *supra* note 56, at 61.

⁵⁹ BORUCHOWITZ, BRINK & DIMINO, *supra* note 12, at 14.

⁶⁰ *Overcriminalization: An Explosion of Federal Criminal Law*, THE HERITAGE FOUND. (Apr. 27, 2011), <http://www.heritage.org/Research/Factsheets/2011/04/OVERCRIMINALIZATION-An-Explosion-of-Federal-Criminal-Law>; Tim Lynch, *Our Overburdened Prison System*, NAT'L REV. ONLINE (Nov. 30, 2009), <http://www.nationalreview.com/corner/190954/our-overburdened-prison-system/tim-lynch>.

⁶¹ AM. BAR ASS'N, CRIMINAL JUSTICE SYSTEM IMPROVEMENTS 1-2 (2008), http://www.americanbar.org/content/dam/aba/migrated/poladv/transition/2008dec_crimjustice.authcheckdam.pdf.

⁶² BORUCHOWITZ, BRINK & DIMINO, *supra* note 12, at 11.

⁶³ *Id.* at 12; *contra* *Argersinger v. Hamlin*, 407 U.S. 25, 38 (1972) ("[T]he prospect of imprisonment for however short a time will seldom be viewed by the accused as a trivial or 'petty' matter and

viduals come into contact with the criminal justice system through traffic and misdemeanor courts,⁶⁴ and significant collateral consequences result from even these seemingly minor convictions.⁶⁵

One consequence of overloaded courts is that misdemeanor defendants frequently lack access to adequate assistance of counsel.⁶⁶ These defendants either do not have legal representation,⁶⁷ or the court assigns a lawyer that is unable to properly represent them⁶⁸ because crushing caseloads reduced most attorneys' interactions with their clients to a "meet-and-plead" relationship.⁶⁹ In this scenario, counsel meet their clients for the first time just before court outside the courtroom, inform their clients of the plea offer, and recommend that it be accepted.⁷⁰ Immediately thereafter, clients enter the plea with no investigation of the circumstances and law surrounding their cases.⁷¹ With the rise of collateral sanctions over the past few decades, one may expect that defense attorneys would spend more time on cases to alleviate the repercussions of a conviction.⁷² Reports fail to show, however, that attorneys are spending more time on misdemeanor cases.⁷³

Admittedly, *Padilla's* significance on the amount of information that attorneys are providing to their clients is unclear. With more attention to collateral consequences given the decision in *Padilla*, academics have opined that defendants will be better informed when deciding whether to accept a plea.⁷⁴ However, *Padilla's* mandate that attorneys address immigration consequences with non-citizens only represents a small percentage of the collateral consequences that arise from convictions.⁷⁵ No other courts have mandated that counsel or the courts have duties to fully and properly

may well result in quite serious repercussions affecting his career and his reputation.") (quoting *Baldwin v. New York*, 399 U.S. 66, 73 (1970)).

⁶⁴ BORUCHOWITZ, BRINK & DIMINO, *supra* note 12, at 11.

⁶⁵ *See supra* Part I.

⁶⁶ BORUCHOWITZ, BRINK & DIMINO, *supra* note 12, at 14.

⁶⁷ *Id.* at 14-15 (explaining that the U.S. court system has become overburdened to the point that nearly one-third of misdemeanor defendants are not even informed of their right to counsel and simply appear in court without the assistance of counsel).

⁶⁸ *Id.* at 14.

⁶⁹ *Id.* at 31.

⁷⁰ *Id.*

⁷¹ *Id.*

⁷² BORUCHOWITZ, BRINK & DIMINO, *supra* note 12, at 34.

⁷³ *See id.* Cf. BUREAU OF JUSTICE ASSISTANCE, OFFICE OF JUSTICE PROGRAMS, U.S. DEP'T OF JUSTICE, KEEPING DEFENDER WORKLOADS MANAGEABLE 4 (2001) (mentioning that felony defendants, facing more serious charges, typically receive more attention from counsel than misdemeanor defendants).

⁷⁴ Chin & Love, *supra* note 56, at 61.

⁷⁵ *See supra* Part I (describing the myriad of collateral consequences that can result from a conviction).

advise defendants who plead guilty of the numerous additional consequences they will face as a result of their convictions.⁷⁶

Further, because collateral consequences are not deemed a direct consequence of a conviction, they are not subject to *ex post facto* challenges.⁷⁷ Therefore, even if defendants have been fully and properly advised of collateral consequences when entering a plea, new or additional sanctions may attach to their convictions in the future because those consequences are not limited to convictions that occurred after the enactment of new legislation. For example, sex offender registries have been challenged—with varying degrees of success—on *ex post facto* grounds by individuals convicted of sex offenses before the relevant registry statutes were passed.⁷⁸ *Padilla*'s narrow holding only provides a defendant with an increase in potential information. *Padilla* does not guarantee an attorney more time to fully investigate a client's case or reduce the impact of a conviction. One could interpret this lack of attention as a violation of a defendant's right to counsel and trial.⁷⁹ Even post-*Padilla*, defendants may suffer the impact of excessive collateral consequences simply because of subsequent laws imposing additional sanctions, counsel's limited time to prepare effective representation, or both.

III. COSTS, BENEFITS, AND EXTERNALITIES OF COLLATERAL CONSEQUENCES

A. *Cost–Benefit Analysis of Select Collateral Consequences*

This section explains the mechanics of a cost–benefit analysis and applies this analysis to the effect of a conviction's collateral consequences on employment, receiving federal student loans, and obtaining housing assistance. A cost–benefit analysis is a useful tool to help objectively evaluate

⁷⁶ Margaret Colgate Love & Gabriel J. Chin, *Padilla v. Kentucky: The Right to Counsel and the Collateral Consequences of Convictions*, CHAMPION, May 2010, available at <http://www.nacdl.org/champion.aspx?id=14611> (noting that *Padilla*'s impact remains unclear with regards to advising defendants of collateral consequences outside of the immigration context).

⁷⁷ U.S. CONST. art. I, § 9, cl. 3; U.S. CONST. art. I, § 10, cl. 1; Jenny Roberts, *The Mythical Divide Between Collateral and Direct Consequences of Criminal Convictions: Involuntary Commitment of "Sexually Violent Predators,"* 93 MINN. L. REV. 670, 678-79 (2008).

⁷⁸ William M. Howard, Annotation, *Validity of State Sex Offender Registration Laws Under Ex Post Facto Prohibitions*, 63 A.L.R. 6TH 351 (2011).

⁷⁹ See U.S. CONST. amend. VI.

economic efficiency.⁸⁰ This analysis typically involves “finding and comparing the costs and benefits of a regulation, tax, or other policy.”⁸¹

In applying this comment’s definition of a cost–benefit analysis to the societal impact of collateral consequences, the concept of social benefit replaces calculating the “hard” value of revenue.⁸² Additionally, the notion of opportunity cost⁸³ replaces calculating monetary cost.⁸⁴ The essential idea behind a cost–benefit analysis is to explain how society will be better off with a different allocation of resources.⁸⁵ “The government’s overall aim is presumably to ensure that social welfare is maximized subject to those constraints over which it has no control such as tastes, technology and resource endowments.”⁸⁶ Careful scrutiny is needed when conducting a cost–benefit analysis in order to value costs and benefits correctly.⁸⁷ While some economic schools of thought, like the Austrian school, maintain that all value is subjective,⁸⁸ this comment adopts the definition of cost as the objective loss of something of value.

In order to properly frame these costs and benefits, it is important to note that a large percentage of those who come in contact with the criminal justice system do so as a result of drug or drug-related offenses. Of the approximately 13,120,947 arrests nationwide in 2010, an estimated 1,638,846—roughly 12%—of arrests were for drug abuse violations.⁸⁹ The overwhelming majority of these arrests were for narcotics possession as opposed to their distribution or manufacture.⁹⁰ Similarly, roughly 20% of inmates in state prisons in 1997—the last year in which data were com-

⁸⁰ HENRY N. BUTLER & CHRISTOPHER R. DRAHOZAL, *ECONOMIC ANALYSIS FOR LAWYERS* 502 (2d ed. 2006).

⁸¹ *Id.* An alternative way of phrasing this analysis is to say that those who gain from some type of change could theoretically compensate those who have lost from the change, and still retain a net gain. *Id.* at 50. Such a change is therefore efficient. *Id.*

⁸² E.J. MISHAN, *COST-BENEFIT ANALYSIS* xxi (3d ed. 1982).

⁸³ Opportunity cost is defined as “[t]he highest valued alternative that must be sacrificed as a result of choosing among alternatives” or “[t]he value placed on opportunities forgone in choosing to produce or consume scarce goods.” BUTLER & DRAHOZAL, *supra* note 80, at 511.

⁸⁴ MISHAN, *supra* note 82.

⁸⁵ *Id.*

⁸⁶ RICHARD LAYARD & STEPHEN GLAISTER, *COST-BENEFIT ANALYSIS* 2-3 (Richard Layard & Stephen Glaister eds., 2d ed. 1994).

⁸⁷ *Id.* at 4.

⁸⁸ Peter J. Boettke, *Austrian School of Economics*, *THE CONCISE ENCYCLOPEDIA OF ECONOMICS*, <http://www.econlib.org/library/Enc/AustrianSchoolofEconomics.html> (last visited Jan. 7, 2012).

⁸⁹ *Crime in the United States 2010: Arrests*, U.S. DEP’T OF JUSTICE & FED. BUREAU OF INVESTIGATION, <http://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s./2010/crime-in-the-u.s.-2010/persons-arrested> (last visited Jan. 7, 2012) (stating that drug abuse violations include the sale, manufacturing, or possession of illicit narcotics).

⁹⁰ *Id.* (highlighting that nationwide, 81.9% of drug abuse violation arrests were for possession of an illicit substance, while only 18.1% of those arrests were for the sale or manufacturing of the same).

piled⁹¹—were serving a sentence for a drug offense.⁹² This figure likely under-represents the impact of drug offenses because it does not include the number of inmates serving sentences for drug-related offenses.⁹³ The annual cost of incarcerating only those state prisoners incarcerated for a drug offense was around \$5 billion.⁹⁴ Of these inmates, nearly 60% have no history of violence or high-level drug activity.⁹⁵

While the nationwide rate of recidivism for drug possession or drug trafficking offenses is not particularly encouraging,⁹⁶ diversion programs such as drug courts⁹⁷ can significantly reduce recidivism.⁹⁸ Given that collateral consequences attach to any criminal conviction, it follows that a significant percentage of those individuals who come into contact with the criminal justice system, regardless of the type of court or the nature of the crime, suffer from the debilitating effect of these consequences. Further, given that alternative punishment programs such as drug courts can effectively rehabilitate former drug users,⁹⁹ the imposition of heavy collateral sanctions on those convicted of drug offenses stymie our criminal justice

⁹¹ *Survey of Inmates in State and Federal Correctional Facilities Resource Guide*, BUREAU OF JUSTICE STATISTICS, <http://www.icpsr.umich.edu/icpsrweb/content/NACJD/guides/sisfcf.html> (last visited Oct. 27, 2010).

⁹² RYAN S. KING & MARC MAUER, THE SENTENCING PROJECT, DISTORTED PRIORITIES: DRUG OFFENDERS IN STATE PRISONS 3 (2002), available at http://www.sentencingproject.org/doc/publications/dp_distortedpriorities.pdf.

⁹³ *Id.* at 3 n.9 (defining a drug-related offense as a crime committed in order to obtain money or other capital to purchase drugs).

⁹⁴ *Id.* at 1.

⁹⁵ *Id.* at 2. U.S. DEP'T OF JUSTICE, AN ANALYSIS OF NON-VIOLENT DRUG OFFENDERS WITH MINIMAL CRIMINAL HISTORIES 6 (1994), available at http://www.fd.org/pdf_lib/1994%20DoJ%20study%20part%201.pdf (stating that high-level drug activity is characterized by an offender who is involved in motivating and/or organizing criminal activity).

⁹⁶ *Prisoner Recidivism Analysis Tool*, BUREAU OF JUSTICE STATISTICS, <http://bjs.ojp.usdoj.gov/index.cfm?ty=datool&url=/recidivism/index.cfm#> (last visited Jan. 7, 2012) (click on "Analysis;" check the "Drug Possession" and "Drug Trafficking" boxes; then click "Generate Results"). Nationwide, one year after release from prison for committing a drug possession or drug trafficking offense, or both, 43.5% of individuals were arrested for a new crime, 26.5% were adjudicated for a new crime, 22.2% were convicted of a new crime, 17.7% were incarcerated for a new crime, and 10.9% were imprisoned for a new crime. *Id.*

⁹⁷ Drug courts are an alternative to the traditional justice system in which substance abuse offenders receive intensive treatment, including frequent random drug screenings and mandatory court appearances to discuss treatment progress. *What are Drug Courts?*, NAT'L ASS'N OF DRUG COURT PROF'LS, <http://www.nadcp.org/learn/what-are-drug-courts> (last visited Dec. 28, 2011); *Types of Drug Courts*, NAT'L ASS'N OF DRUG COURT PROF'LS, <http://www.nadcp.org/learn/what-are-drug-courts/types-drug-courts> (last visited Dec. 28, 2011).

⁹⁸ JOHN ROMAN ET AL., U.S. DEP'T OF JUSTICE, RECIDIVISM RATES FOR DRUG COURT GRADUATES: NATIONALLY BASED ESTIMATES, FINAL REPORT 2 (2003), available at <https://www.ncjrs.gov/pdffiles1/201229.pdf> (explaining that, nationwide, one year after graduating from a drug court program, only 16.4% of graduates were arrested and charged with an offense that carried a sentence of at least one year in prison).

⁹⁹ *Id.*

system's utilitarian goals of rehabilitation and reintegration.¹⁰⁰ As this comment demonstrates, it appears that the cost of imposing collateral consequences outweighs the benefit of rehabilitation.

Employing a cost-benefit analysis to evaluate the economic efficiency of barring ex-offenders from certain types of employment demonstrates that the costs of such a ban outweigh the benefits. Specifically, one collateral consequence of a misdemeanor or felony conviction is the lost opportunity of federal employment. There are a multitude of crimes for which these individuals can be barred from a federal job, either permanently or temporarily.¹⁰¹ The removal of ex-offenders from federal office or employment arguably benefits society. The rationale in support of this ban is the belief that ex-convicts cannot be trusted with or are less deserving of this type of employment.¹⁰² Therefore, society benefits by keeping those with a criminal conviction on their record from violating the duties and responsibilities of these important jobs.

However, excluding allegedly untrustworthy individuals from federal office or employment comes with significant punitive costs for both the individual and society. As a whole, offenders are disadvantaged in terms of employment even before their initial arrest due to sporadic employment and lack of education.¹⁰³ Data from 1997 indicate that only about two-thirds of all prisoners were employed before their arrest,¹⁰⁴ and just over one-third of these prisoners had graduated from high school.¹⁰⁵ The immediate cost of this collateral consequence of conviction is that these individuals are barred from employment with the federal government, which employs millions of individuals¹⁰⁶ and encompasses positions from the managerial level to in-

¹⁰⁰ KENT GREENAWALT, *Punishment* (1999), reprinted in JOSHUA DRESSLER, *CASES AND MATERIALS ON CRIMINAL LAW* 34-36 (5th ed. 2009).

¹⁰¹ Such crimes include treason (18 U.S.C. § 2381 (2006)), bribing a public official (*id.* § 201(b)), inciting a riot or civil disorder (5 U.S.C. § 7313 (2006)), attempting to overthrow the federal government (18 U.S.C. §§ 2385, 2387 (2006)), trading in public funds or property (*id.* § 1901), unlawfully disclosing business trade secrets (*id.* § 1905), using federal money to finance the lobbying of a member of Congress (*id.* § 1913), concealing or mutilating public documents (*id.* § 2071), unlawfully disclosing or inspecting taxpayer return information (26 U.S.C. §§ 7213(a)(1), (b), 7213A(b)(2) (2006)), or committing extortion, bribery, or conspiracy to defraud the United States (*id.* § 7214(a)).

¹⁰² Demleitner, *supra* note 37, at 161.

¹⁰³ Christopher Stafford, Note, *Finding Work: How to Approach the Intersection of Prisoner Reentry, Employment, and Recidivism*, 13 *GEO. J. ON POVERTY L. & POL'Y* 261, 262-63 (2006).

¹⁰⁴ *Id.* at 262.

¹⁰⁵ *Id.* at 263.

¹⁰⁶ Telephone Interview with Colleen Teixeira Moffat, Economist, Emp't Projections Program, Bureau of Labor Statistics (July 17, 2012) (stating that the industry of the federal government represents 2.1% of employment in the United States, or 2.9 million jobs); BILL HEBENTON & TERRY THOMAS, *CRIMINAL RECORDS: STATE, CITIZEN AND THE POLITICS OF PROTECTION* 111 (1993) (demonstrating that federal and state bans on public employment removes ex-offenders from eligibility for 350 occupations, composed of some ten million jobs).

stallation, maintenance, and repair staff.¹⁰⁷ While such a ban may seem appropriate, it is not even required that the ex-offender's offense be related to the employment from which the ex-offender is subsequently barred.¹⁰⁸ Shutting off an enormous source of jobs necessarily increases the difficulty this population faces in their search for gainful employment.¹⁰⁹

This difficulty is heightened during tough economic times when jobs are scarce. Absent a period of recession, the removal of this segment of the population from federal employment detracts from the tax base of a given community. Even though wealth is created in the private sector, ex-offenders are also routinely barred from a wide swath of private sector jobs due to licensing restrictions.¹¹⁰ Therefore, the additional removal of ex-offenders from public sector jobs further impacts the ex-offender's community by prohibiting that individual from contributing to the tax base. As fewer income, Social Security, and Medicare taxes are collected, state governments and the federal government have less money available to maintain services shared by all community members. The drain of unemployment on a community demonstrates that the costs of barring individuals who have faced a criminal conviction from federal jobs exceed the benefits, as those precluded from lawful employment in either the public or private sector frequently commit new crimes to make a living.¹¹¹ Rather than a safety-oriented consequence, the broad ban on employment only serves a punitive function that harms not only the individual, but his or her community as well. A simple way to diminish this collateral consequence would be to bar ex-offenders only from those jobs specifically related to their criminal conviction.¹¹² Another solution would be to allow individuals to petition the

¹⁰⁷ Telephone Interview with Colleen Teixeira Moffat, *supra* note 106 (stating that within the federal government, 6% of jobs are within the management occupation and 4% of jobs are within the installation, maintenance, and repair occupations).

¹⁰⁸ Michael Pinard & Anthony C. Thompson, *Offender Reentry and the Collateral Consequence of Criminal Convictions: An Introduction*, 30 N.Y.U. REV. L. & SOC. CHANGE 585, 596 (2006) (citing Nora V. Demleitner, "Collateral Damage": No Re-entry for Drug Offenders, 47 VILL. L. REV. 1027, 1038-39 (2002)).

¹⁰⁹ *Id.* at 596-97 (citing Nora V. Demleitner, "Collateral Damage": No Re-entry for Drug Offenders, 47 VILL. L. REV. 1027, 1038 (2002)).

¹¹⁰ *Id.* at 597 (citing Bruce E. May, *The Character Component of Occupational Licensing Laws: A Continuing Barrier to the Ex-Felon's Employment Opportunities*, 71 N.D. L. REV. 187, 194-95 (1995)) (explaining that nearly two thousand state statutory provisions exist, and, depending on the state, bar ex-offenders from obtaining licenses to become—among other things—barbers, beauticians, and nurses).

¹¹¹ Josephine R. Potuto, *A Model Proposal to Avoid Ex-Offender Employment Discrimination*, 41 OHIO ST. L.J. 77, 81 (1980) (noting that to reduce recidivism, ex-offenders should have access to a wide variety of employment choices).

¹¹² Clark, *supra* note 7, at 205. Certain professions follow this standard. For example, an attorney, as an officer of the court, can lose the ability to practice law if she is convicted of a crime. MODEL FED. RULES OF DISCIPLINARY ENFORCEMENT R. 1 (1991), *available at* <http://www.americanbar.org/content/dam/aba/migrated/cpr/discipline/mfrde.authcheckdam.pdf>. Additionally, doctors can lose the ability to practice medicine if their conduct threatens the welfare or safety

court in which a conviction was obtained for the expungement of a first-time conviction.¹¹³

An additional example of the costs of collateral consequences outweighing their benefits involves the exclusion of those individuals convicted of possessing or selling drugs from receiving federal education assistance in the form of student loans.¹¹⁴ The rationale for this collateral consequence is that these individuals have demonstrated their untrustworthiness by failing to follow the laws.¹¹⁵ Therefore, society benefits by removing the temptation for ex-offenders to abuse student loan programs. However, those students excluded from federal loan programs lose access to a significant financial resource that the majority of undergraduate students rely on to finance their education.¹¹⁶ Without access to this source of funding, it becomes economically unrealistic for many ex-offenders to attend college.¹¹⁷

The ramifications of not having a college degree in today's economy are significant, as college graduates can expect to earn up to \$300,000 more than non-college graduates over a typical forty-year career.¹¹⁸ Additionally, a college degree can open the door to advanced degrees and often leads to higher quality jobs that include benefits like health insurance.¹¹⁹ Those convicted of a crime are barred from these jobs because they cannot afford to attend college in the first place. Removing access to student loans precludes female ex-offenders from joining in with the trend of women now earning more advanced degrees than men.¹²⁰ Further, people aged twenty-nine and under are convicted of drug offenses at a higher rate than those

of the public. *Revocation and Suspension of Physician Licenses*, USLEGAL.COM, <http://physicians.uslegal.com/revocation-and-suspension-of-physician-licenses> (last visited January 3, 2013). Finally, in some states a nurse's license can automatically be suspended for felony drug convictions. See, e.g., ALA. CODE § 34-21-25(b)(1)(f) (2011); 63 PA. CONS. STAT. ANN. § 422.40 (West 1986).

¹¹³ See *infra* Part V.

¹¹⁴ 20 U.S.C. § 1091(r) (2006).

¹¹⁵ Demleitner, *supra* note 37, at 161.

¹¹⁶ Christina Chang Wei et al., *Web Tables: Undergraduate Financial Aid Estimates by Type of Institution in 2007-08*, U.S. DEP'T OF EDUC., NAT'L CTR. FOR EDUC. STATISTICS, Table 1 (2009), <http://nces.ed.gov/pubs2009/2009201.pdf>.

¹¹⁷ Demleitner, *supra* note 37, at 158 (noting that the denial of social and welfare rights is economically disadvantageous to ex-offenders).

¹¹⁸ Kim Clark, *How Much is that College Degree Really Worth?*, U.S. NEWS & WORLD REP. (Oct. 30, 2008), <http://www.usnews.com/education/articles/2008/10/30/how-much-is-that-college-degree-really-worth>. This seemingly low number factors in inflation and student loan debt. Unadjusted, the typical college graduate can expect to earn \$800,000 more than the non-college graduate.

¹¹⁹ *Id.*

¹²⁰ U.S. DEP'T OF EDUC., NAT'L CTR. FOR EDUC. STATISTICS, *THE CONDITION OF EDUCATION 2011 Table A-26-2* (2011), available at <http://nces.ed.gov/fastfacts/display.asp?id=72>.

aged thirty and over.¹²¹ These same young people are also the individuals seeking student loans in order to better their lives after a conviction.¹²² Banning young people convicted of a drug offense from receiving student loans—which often removes the opportunity to attend college—precludes those individuals from obtaining higher paying jobs and contributing to their communities at a higher level during their working years. These individuals are instead diverted into lower paying jobs. This relegation is a punitive rather than rehabilitative sanction that perpetuates a permanent cycle of underemployment. Here again, the costs of banning ex-offenders from obtaining loans for college outweigh the potential benefits of the regulation.¹²³

As a final example, the cost of removing ex-offenders from receiving government housing assistance outweighs its benefit. Statutory provisions providing for federally subsidized housing require that tenants remain drug-free and disqualify lawbreakers.¹²⁴ Therefore, it is arguable that ex-offenders who have demonstrated a propensity to violate the law should not be trusted to abide by these provisions and therefore should be banned from

¹²¹ Between 1993 and 2001, there was an average of 20,131.7 arrests per 100,000 inhabitants for drug abuse violations for those individuals 29 and under, while the average arrest rate for the same crimes was only 3,124.2 for those individuals 30 and over. FED. BUREAU OF INVESTIGATION UNIF. CRIME REPORTING PROGRAM, AGE-SPECIFIC ARREST RATES AND RACE-SPECIFIC ARREST RATES FOR SELECTED OFFENSES, 1993-2001 41-42 (2003), available at http://www.fbi.gov/about-us/cjis/ucr/additional-ucr-publications/age_race_arrest93-01.pdf (these averages were calculated by adding up the arrest rates for each year for the respective age groups and then dividing by the nine-year time span that the report covers). Further, the average age of those arrested for drug abuse violations, which includes the sale, use, growing and manufacturing of narcotic drugs, each year between 1993 and 2001 was between 28 and 29 years old. *Id.* at 49, 51.

¹²² Among the undergraduate population, those students who do not take out private loans are, on average, 26.6 years old, while those students who do take out private loans average 23.5 years. AM. COUNCIL ON EDUC., WHO BORROWS PRIVATE LOANS? 3 (2007), available at <http://www.acenet.edu/AM/Template.cfm?Template=/CM/ContentDisplay.cfm&ContentID=23410>. While students aged 38 to 41 had the greatest increase in student loan debt between 2008 and 2011, students between ages 26 and 29 still have the highest average level of student loan debt. Mitch Lipka, *Middle-Aged Borrows Piling on Student Debt*, REUTERS (Dec. 27, 2011), <http://www.reuters.com/article/2011/12/27/us-studentdebt-middleage-idUSTRE7BQ0T620111227>.

¹²³ It could be argued, however, that the cost of banning ex-offenders from receiving student loans is justified in light of the growing federal deficit and the increase in students defaulting on their loans. Kevin Helliker, *Student Loan Defaults on Rise*, WALL ST. J. (Sept. 13, 2011), <http://online.wsj.com/article/SB10001424053111904353504576566791840707646.html> (describing sharp increase in recent years in students defaulting on their loans). For example, due to deficit concerns, Congress recently approved a measure ending federal subsidies of graduate student loans. Zaid Jilani, *Debt Deal Would End Subsidized Loans to Grad Students, Produce Savings Equal to Only Three Months in Afghanistan*, THINK PROGRESS (Aug. 1, 2011), <http://thinkprogress.org/education/2011/08/01/284804/subsidized-loans-afghanistan/>. However, if deficit and default considerations are to be taken into account in reducing federal subsidies for loans, all students, rather than only those with criminal histories, should be equally affected.

¹²⁴ 42 U.S.C. §§ 13661, 1437f(d)(1)(B)(iii) (2006).

receiving housing assistance. However, lack of housing combined with lack of employment compound one another,¹²⁵ straining the impoverished communities that this population calls home.¹²⁶ Without access to proper housing, ex-offenders frequently recidivate, rather than reintegrate into the community.¹²⁷

This punitive sanction perpetuates the persistence of crime, precluding economic improvement for the ex-offender's community. High crime rates in a given neighborhood tend to lead to families and individuals either moving out of an area or not settling there in the first place, ultimately leading to a less consistent population.¹²⁸ Without a stable population in a community, there is less predictability in what types of public services will be provided and in which sector private enterprises will settle in a given area, thereby decreasing the probability that the community will amass wealth.¹²⁹ The median daily cost to taxpayers of providing community housing to an ex-offender is \$30.48, which is significantly lower than that same measurement of the cost of housing an individual in prison (\$59.43) or in jail (\$70.00).¹³⁰ Therefore, the benefits from the aggregate wealth of low-crime neighborhoods likely outweigh the much lower costs of providing housing to those convicted of a crime. As such, statutory provisions that prevent ex-offenders from obtaining housing should be modified to allow ex-offenders to obtain housing assistance.

B. *Negative Externalities of Collateral Consequences*

Not all economic decision-makers bear the costs and benefits of their actions.¹³¹ Sometimes, costs and benefits are passed along to people who are not parties to a transaction. These third-party effects are called externalities.¹³² Neither positive nor negative externalities are socially desirable. Free markets aim to force individuals to internalize the costs and benefits of their decisions in order to optimize efficiency.¹³³ Negative externalities

¹²⁵ Pinard & Thompson, *supra* note 108, at 595.

¹²⁶ Jeremy Travis, Laurie O. Robinson & Amy L. Solomon, *Prisoner Reentry: Issues for Practice and Policy*, CRIM. JUST., Spring 2002, at 12.

¹²⁷ *Id.* at 14.

¹²⁸ David S. Kirk & John H. Laub, *Neighborhood Change and Crime in the Modern Metropolis*, 39 CRIME & JUST. 441, 457-59 (2010).

¹²⁹ William H. Frey, *Central City White Flight: Racial and Nonracial Causes*, 44 AM. SOC. REV. 425, 427-28 (1979).

¹³⁰ KATHERINE CORTES & SHAWN ROGERS, COUNCIL OF STATE GOV'TS JUSTICE CTR., REENTRY HOUSING OPTIONS: THE POLICYMAKERS' GUIDE viii (2010), available at http://www.ojp.usdoj.gov/BJA/pdf/CSG_Reentry_Housing.pdf.

¹³¹ BUTLER & DRAHOZAL, *supra* note 80, at 175.

¹³² *Id.*

¹³³ Tony Cleaver, *ECONOMICS: THE BASICS* 226-27 (2d ed. 2011).

occur when “the social costs of producing or consuming a good are greater than the private costs.”¹³⁴

After completing a prison sentence, released individuals disproportionately return to poor urban areas, further burdening these struggling communities with a demand for jobs and services that they are unable to handle.¹³⁵ The collateral consequences that attach to one’s conviction introduce significant negative externalities that prevent depressed local economies from recovering. Like any other community member, ex-offenders require community-based services¹³⁶ such as substance abuse treatment and mental health counseling. However, these individuals are barred from a myriad of jobs and from access to student loans that would allow them to return to school to obtain an education.¹³⁷ Hence a negative externality results, as these individuals use community services without contributing to the tax base that maintains that service. Former inmates, stymied by a lack of employment opportunities and access to services, often return to criminal behaviors to earn money¹³⁸ rather than contribute in a lawful manner to the community.¹³⁹ Therefore, collateral consequences generate significant negative externalities for an ex-offender’s surrounding community.

Further, those convicted of a crime can also lose custody of their children.¹⁴⁰ This deprivation of custody creates negative externalities for society, as the state or other family members must then take over care of dependent children. Not surprisingly, society as a whole suffers disproportionately from an increase in female incarceration.¹⁴¹ In 2000, it cost the public an estimated \$25,000 per year to house an incarcerated female and an additional \$25,000 for foster care for each of her dependent children.¹⁴² The loss of custody that results from a conviction creates burdens for the state

¹³⁴ BUTLER & DRAHOZAL, *supra* note 80, at 511.

¹³⁵ Pinard, *supra* note 14, at 628.

¹³⁶ Deborah N. Archer & Kele S. Williams, *Making America “The Land of Second Chances”*: *Restoring Socioeconomic Rights for Ex-Offenders*, 30 N.Y.U. REV. L. & SOC. CHANGE 527, 529 (2006).

¹³⁷ *See supra* Part III.A.

¹³⁸ Potuto, *supra* note 111, at 81-82 (noting that “lack of employment increases the chances that the ex-offender will recidivate.”).

¹³⁹ Over two-thirds of prisoners arrested in 1994 were arrested within three years for a new offense. PATRICK A. LANGAN & DAVID J. LEVIN, U.S. DEP’T OF JUSTICE, *RECIDIVISM OF PRISONERS RELEASED IN 1994*, at 1 (2002), available at <http://bjs.ojp.usdoj.gov/content/pub/pdf/rpr94.pdf>.

¹⁴⁰ “Under the federal Adoption and Safe Families Act (ASFA), whenever a child has lived in foster care for 15 of the most recent 22 months, the state is required to file a petition to terminate parental rights. Although the median minimum sentence for a female offender is 36 months, ASFA makes no exception for incarcerated parents.” *Words from Prison – Did You Know...?*, AM. CIV. LIBERTIES UNION (June 12, 2006), <http://www.aclu.org/womens-rights/words-prison-did-you-know>.

¹⁴¹ “Women are the fastest growing segment of the incarcerated population.” *Id.*

¹⁴² LENORA LAPIDUS ET AL., AM. CIVIL LIBERTIES UNION, *CAUGHT IN THE NET: THE IMPACT OF DRUG POLICIES ON WOMEN AND FAMILIES* 19 (2005), available at http://www.aclu.org/files/images/asset_upload_file431_23513.pdf.

and private individuals in terms of child care that parents, especially mothers, would more often be able to internalize absent such consequences.

IV. FEMALE EX-OFFENDERS AND COLLATERAL CONSEQUENCES: AN INQUIRY

Women are still overwhelmingly the primary caretakers of dependent children.¹⁴³ The number of women in the criminal justice system continues to pale in comparison to men,¹⁴⁴ as women comprised only 6.6% of the total number of inmates in state and federal prisons in 2000.¹⁴⁵ However, the female inmate population has skyrocketed over the past two decades, growing at nearly twice the rate of men.¹⁴⁶ The rate of growth of incarcerated mothers over the past two decades outpaced the rate of growth of incarcerated fathers.¹⁴⁷ The number of mothers incarcerated from 1991 to 2007 increased 122%, while the number of fathers incarcerated during that same period increased at a much lower rate of 76%.¹⁴⁸ Even though the number of female inmates falls far behind the number of male inmates, the increased rate of growth of incarcerated mothers compared to fathers suggests that many collateral consequences nevertheless disproportionately impact women.

Women, incarcerated or not, overwhelmingly remain responsible for child care duties. Female relatives assume responsibility over dependent children when mothers are incarcerated,¹⁴⁹ and ex-offender mothers are the parent regaining or attempting to regain custody of their children once they are released back into their communities.¹⁵⁰ When searching for housing

¹⁴³ NEW AM. FOUND., THE WAY WOMEN WORK (Mar. 2004), http://www.newamerica.net/files/archive/Doc_File_1504_1.pdf.

¹⁴⁴ Myrna S. Raeder, *A Primer on Gender-Related Issues That Affect Female Offenders*, CRIM. JUST., Spring 2005, at 4.

¹⁴⁵ ALLEN J. BECK & PAIGE M. HARRISON, U.S. DEP'T OF JUSTICE, PRISONERS IN 2000 5 (2001), available at <http://bjs.ojp.usdoj.gov/content/pub/pdf/p00.pdf>.

¹⁴⁶ LAWRENCE A. GREENFELD & TRACY L. SNELL, U.S. DEP'T OF JUSTICE, WOMEN OFFENDERS 6 (1999), available at <http://bjs.ojp.usdoj.gov/content/pub/pdf/wo.pdf> ("The number of women per capita involved in corrections overall has grown 48% since 1990, compared to a 27% increase in the number of men per capita."); Raeder, *supra* note 144, at 4.

¹⁴⁷ LAUREN E. GLAZE & LAURA M. MARUSCHAK, U.S. DEP'T OF JUSTICE, PARENTS IN PRISON AND THEIR MINOR CHILDREN 2 (2010), available at <http://bjs.ojp.usdoj.gov/content/pub/pdf/pptmc.pdf>.

¹⁴⁸ *Id.*

¹⁴⁹ Forty-two percent of incarcerated mothers reported leaving their dependent children with the child's grandmother. INST. ON WOMEN & CRIMINAL JUSTICE, INCARCERATED MOTHERS AND THEIR CHILDREN: HIGHLIGHTS FROM THE NEW FEDERAL REPORT (2008), http://www.wpaonline.org/pdf/2008_BJS_parents_Final.pdf (citing GLAZE & MARUSCHAK, *supra* note 147, at 5).

¹⁵⁰ Only 37% of incarcerated mothers reported that the child's father was the child's caregiver while she was incarcerated. GLAZE & MARUSCHAK, *supra* note 147, at 5. Because of the lack of in-

after her release, a mother must face the reality that individuals convicted of any “criminal activity that threatens the health, safety, or right to peaceful enjoyment of the premises by other tenant[s]” can be evicted from federally funded public housing.¹⁵¹ The potential for eviction is applicable regardless of whether such a threat stems from the tenant herself or a member of her household.¹⁵² Therefore, women convicted of a crime have a hard time finding housing given the statute’s restrictions. Because the statute also removes assistance to those who have an ex-offender living in their household, women with criminal convictions on their record cannot readily rely on friends and family who are living in federally subsidized housing. Those female ex-offenders who cannot secure housing suffer from a lack of home and job stability, as do their dependent children. Further, this bar from federally funded public housing typically arises as a result of a drug conviction, and the rate of increase in drug offense convictions is higher for women than it is for men.¹⁵³

Because women are more likely than men to be the caretakers of children prior to incarceration,¹⁵⁴ loss of housing assistance as a consequence of their conviction places a heavier burden on female ex-offenders who return to these same childcare duties. Without access to more affordable housing, these women must divert more of their limited resources to housing that can accommodate both themselves and their dependent children. Male ex-offenders, who do not have to bear the same burden of childcare, are not similarly disadvantaged.

Those convicted of a crime may also be evicted from federally funded housing for violating a condition of probation or parole.¹⁵⁵ These conditions often limit the geographic area in which these individuals can travel.¹⁵⁶

volvement of a child’s father, the children of incarcerated mothers are much more likely to enter the foster care system. *Id.* Therefore, female ex-offenders are the parent focused on regaining custody of their children when they are released. Anthony C. Thompson, *Navigating the Hidden Obstacles to Ex-Offender Reentry*, 45 B.C. L. REV. 255, 284 (2004).

¹⁵¹ 42 U.S.C. § 1437f(d)(1)(B)(iii) (2006) (emphasizing that such criminal activity includes drug-related crimes, physical violence, domestic violence, dating violence, and stalking).

¹⁵² *Id.*

¹⁵³ The increase in inmates during the 1990s for drug offenses among women was 35%, while the rate of increase for men was only 19%. BECK & HARRISON, *supra* note 145, at 12 tbl.17. Additionally, between 1980 and 2009, the male arrest rate for drug possession or use doubled while the female arrest rate for the same offenses tripled. HOWARD N. SNYDER, U.S. DEP’T OF JUSTICE, ARREST IN THE UNITED STATES, 1980-2009 12 (2011), available at <http://www.bjs.gov/content/pub/pdf/aus8009.pdf>.

¹⁵⁴ “Prior to incarceration, women were more likely than [sic] men to live with their children, be the primary caregiver and to be the head of a single-parent household. In the month prior to arrest or immediately before incarceration, 64.2% of mothers lived with their child(ren) compared to 46.5% of fathers.” INST. ON WOMEN & CRIMINAL JUSTICE, *supra* note 149 (citing GLAZE & MARUSCHAK, *supra* note 147, at 4-5).

¹⁵⁵ 42 U.S.C. § 1437f(d)(1)(B)(v)(II) (2006).

¹⁵⁶ *Gall v. United States*, 552 U.S. 38, 48-49 (2007) (describing the typical conditions of probation).

Given the employment restrictions that ex-offenders face, it often becomes necessary to travel outside of these bounds to find adequate paid employment. These restrictions burden women more heavily than men because women continue to face employment discrimination.¹⁵⁷ Holding other factors such as education and type of work constant, a woman still earns only eighty cents for every dollar that a man earns.¹⁵⁸ Female employees have gained ground in certain segments of the workforce, such as leisure and hospitality, and financial activities.¹⁵⁹ However, women remain woefully underrepresented in industries like information, construction, and manufacturing.¹⁶⁰ Given the combined effects of a conviction's repercussions and the employment discrimination that women still face, it is noticeable that women are disproportionately burdened by the collateral consequences of a criminal conviction.

Further, individuals convicted of a federal or state felony for the possession, use, or distribution of a controlled substance are not eligible to receive food stamps or temporary assistance to needy families.¹⁶¹ The amount of this federal aid that would have been paid to the family is reduced by the amount that an ex-offender in the household would otherwise have received.¹⁶² Women disproportionately seek these federal benefits due to child care obligations.¹⁶³ Women who have been convicted of a crime therefore are more significantly burdened by this ban once they resume their childcare obligations.

This disproportionate impact of collateral consequences on women extends further to negatively impact their dependent children. Children are negatively impacted by a parent's incarceration in many ways.¹⁶⁴ A parent's incarceration is detrimental to a child's social capital, or "the ability of actors to secure benefits by virtue of membership in social networks or other social structures."¹⁶⁵ In other words, children who do not have stability

¹⁵⁷ Ruth Mayhew, Hiring Discrimination Against Women, CHRON.COM, <http://smallbusiness.chron.com/hiring-discrimination-against-women-2861.html> (last visited Jan. 7, 2012).

¹⁵⁸ U.S. CONGRESS JOINT ECON. COMM., WOMEN AND THE ECONOMY 2010: 25 YEARS OF PROGRESS BUT CHALLENGES REMAIN 1 (2010), available at [http://op.bna.com/dlrcases.nsf/id/lswr-88nlnb/\\$File/Women%20Workers%20JEC.pdf](http://op.bna.com/dlrcases.nsf/id/lswr-88nlnb/$File/Women%20Workers%20JEC.pdf).

¹⁵⁹ *Id.* at 4, 5 fig.5.

¹⁶⁰ *Id.*

¹⁶¹ 21 U.S.C. § 862a(a) (2006).

¹⁶² *Id.* § 862a(b).

¹⁶³ "The ban is devastating for women who need cash assistance to help support their families immediately after release, especially because 30% of women in prison were on welfare in the month prior to their arrest." AM. CIV. LIBERTIES UNION, *supra* note 140 (citing GREENFELD & SNELL, *supra* note 146, at 8).

¹⁶⁴ John Hagan & Ronit Dinovitzer, *Collateral Consequences of Imprisonment for Children, Communities, and Prisoners*, 26 CRIME & JUST. 121, 123 (1999).

¹⁶⁵ Alejandro Portes, *Social Capital: Its Origins and Applications in Modern Sociology*, 24 ANN. REV. SOC. 1, 6 (1998).

in their home life lack the ability to develop bonds within their community. The disruption in family life caused by incarceration leads to “the strains of economic deprivation, the loss of parental socialization through role modeling, support, and supervision, and the stigma and shame of societal labeling.”¹⁶⁶ Additionally, once mothers get out of prison, they may not be prepared to return to the rigors of motherhood, especially considering the difficulty they face in obtaining housing, employment, and child care.¹⁶⁷ Emotionally and fiscally unprepared mothers, suffering the impact of collateral consequences, are frequently unable to provide a stable and happy home for their dependent children.¹⁶⁸ Because women are more often the caretakers of young children, and because certain child-focused collateral consequences disproportionately impact women, the dependent children of these women are likely disproportionately impacted as well. This perpetuates emotional and economic problems that can permanently trap young children in a cycle of poverty.

Collateral consequences have a disproportionate effect on female ex-offenders, as compared to males. Especially when viewing these consequences in conjunction with one another, female ex-offenders face increased difficulty in reintegrating into their communities because of the barriers that collateral consequences create. Because of this disproportionate effect, collateral consequences indirectly discriminate against women even though they are incarcerated at a far lower rate than men.

V. REDUCING THE COST OF COLLATERAL CONSEQUENCES THROUGH DECRIMINALIZATION, EXPUNGEMENT, AND EX POST FACTO CONSIDERATIONS

Many behaviors that were once seen as simply socially unacceptable have today been criminalized.¹⁶⁹ However, simply because the majority of society does not condone a particular behavior, it does not necessarily follow that such a behavior should be burdened with the stigma of criminal sanctions. In a culturally diverse society, certain behaviors will not receive universal acceptance. The federal and state governments have tended to criminalize many of these behaviors rather than imposing other methods of regulation, such as fines.¹⁷⁰ Examples of increased criminalization of cer-

¹⁶⁶ Hagan & Dinovitzer, *supra* note 164, at 123.

¹⁶⁷ *Id.* at 144.

¹⁶⁸ *Id.*

¹⁶⁹ BORUCHOWITZ, BRINK & DIMINO, *supra* note 12, at 25.

¹⁷⁰ See Michael N. Giuliano, *The “Risk” of Liberty: Criminal Law in the Welfare State*, THE FREEMAN, Sept. 2008, <http://www.thefreemanonline.org/featured/the-risk-of-liberty-criminal-law-in-the-welfare-state/> (describing the criminalization of behaviors now considered negligent in several professional settings, including management, bartending, and medicine); Steven J. Tepper, *Stop the Beat: Quiet Regulation and Cultural Conflict*, 24 SOC. F. 276, 279-81 (2009) (detailing gov-

tain behaviors through legislation include smoking in public places¹⁷¹ and the vacillation in the legal drinking age.¹⁷² Given this trend of over-criminalization, courts' misdemeanor dockets are clogged with cases that most defense attorneys feel should not warrant jail time.¹⁷³

To ameliorate the negative impact of collateral consequences, legislatures should decriminalize those offenses that do not significantly threaten public safety.¹⁷⁴ Examples of such efforts include: Hawaii's move to decriminalize crimes related to agriculture, conservation, transportation, and boating; Massachusetts's effort to decriminalize drug possession; and Nebraska's endeavor to decriminalize dog leash and trespass offenses.¹⁷⁵ By removing the criminal sanction of a misdemeanor conviction and instead reclassifying these behaviors into status offenses, penalties, or infractions, the collateral consequences that once attached to the criminal sanction would be significantly reduced.¹⁷⁶ Further, the government would still have the means available to incentivize certain behaviors seen as more socially desirable. Fines and sanctions would serve to deter individuals from engaging in those behaviors viewed as undesirable by a majority of the population.¹⁷⁷ Without the impact of a misdemeanor conviction, far fewer individuals would have to cope with the repercussions of collateral consequences.

ernment crackdowns on people that listened to jazz; read comic books; listened to rock and roll, punk, and rap music; and, most recently, "quiet regulations" aimed at ending youth-oriented raves).

¹⁷¹ State smoking bans have been on the rise in recent years. Patti Neighmond, *Smoking Bans Help People Quit, Research Shows*, NAT'L PUB. RADIO (Oct. 25, 2007), <http://www.npr.org/templates/story/story.php?storyId=15610995>. State statutory provisions that ban smoking in public places include: ALA. CODE § 22-15A-4 (2011), ARIZ. REV. STAT. ANN. § 36-601.01 (2011), COLO. REV. STAT. ANN. § 25-14-204 (West 2011), D.C. CODE § 7-1703 (2011), KAN. STAT. ANN. § 21-6110 (West 2011), N.Y. PUB. HEALTH LAW § 1399-o (McKinney 2011), OHIO REV. CODE ANN. § 3791.031 (West 2011), TENN. CODE ANN. § 39-17-1803 (West 2011), and VA. CODE ANN. § 15.2-2824 (West 2011).

¹⁷² The minimum legal drinking age was set at twenty-one after Prohibition, was changed by twenty-nine states to eighteen, nineteen, or twenty between 1970 and 1975, and was again set at twenty-one in 1984 by the federal government. *Minimum Legal Drinking Age*, AM. MED. ASS'N, <http://www.ama-assn.org/ama/pub/physician-resources/public-health/promoting-healthy-lifestyles/alcohol-other-drug-abuse/facts-about-youth-alcohol/minimum-legal-drinking-age.page> (last visited Jan. 4, 2012).

¹⁷³ BORUCHOWITZ, BRINK & DIMINO, *supra* note 12, at 25-26.

¹⁷⁴ *Id.* at 27.

¹⁷⁵ *Id.*

¹⁷⁶ *Id.* at 28. While many collateral consequences are necessarily eliminated by the reclassification of formerly criminal behaviors to offenses, consequences such as deportation can attach to those drug offenses that a state defines as a noncriminal drug offense, like the possession of a small amount of marijuana. See Roberts, *supra* note 77, at 674 (citing 8 U.S.C. § 1227(a)(2)(B)(i) (2006) and N.Y. PENAL LAW § 221.05 (McKinney 2011), which makes unlawful possession of marijuana a noncriminal offense).

¹⁷⁷ Gary S. Becker, *Crime and Punishment: An Economic Approach*, 76 J. OF POL. ECON. 169, 193-98 (1968) (arguing that fines as a method of punishment should be used whenever feasible in order to increase social welfare by achieving the optimal level of crime).

Another simple suggestion for ameliorating the harsh effects of collateral consequences would be to tie such consequences to *ex post facto*¹⁷⁸ considerations.¹⁷⁹ Defendants would necessarily be able to make better informed pleas if legislatures were prevented from adding collateral sanctions to criminal convictions once a defendant has been found guilty.¹⁸⁰ Given that *ex post facto* considerations cannot be adequately considered at the time an accused decides whether or not to plead guilty to an offense, such consequences become unnecessarily punitive.

One example of an unexpected and significant negative consequence of over-criminalization is the offense of driving on a suspended license. The cycle of this offense often begins innocuously, with an individual receiving a minor traffic infraction and a resulting fine.¹⁸¹ If the individual fails to pay that fine, her license is suspended.¹⁸² However, because the majority of Americans outside of metropolitan areas lack easy access to public transportation, the person charged usually cannot give up driving, given basic needs for transportation such as commuting to work and buying groceries.¹⁸³ Subsequent driving could result in the criminal charge of driving with a suspended license, exposing that individual to additional fines, suspension of driving privileges, or even incarceration.¹⁸⁴ The issue compounds upon itself, leading to a cycle of debt that an individual cannot escape for what most people would consider a very minor crime.¹⁸⁵ Decriminalization of this offense would remove one instance of collateral conse-

¹⁷⁸ U.S. CONST. art. I, § 9, cl. 3; U.S. CONST. art. I, § 10, cl. 1.

¹⁷⁹ The Constitution's *ex post facto* clause "prevents the state and federal governments from passing laws that have a retroactive effect, but it only applies to statutes that increase criminal punishment for crimes that occurred before the passage of the statute." Brian Kleinhaus, Note, *Serving Two Masters: Evaluating the Criminal or Civil Nature of the VWPA and MVRA Through the Lens of the Ex Post Facto Clause, the Abatement Doctrine, and the Sixth Amendment*, 73 *FORDHAM L. REV.* 2711, 2713 (2005) (citing *Calder v. Bull*, 3 U.S. 386, 390 (1798)).

¹⁸⁰ For example, *ex post facto* considerations with regards to the Victims Witness and Protection Act's (VWPA) and the Mandatory Victim Restitution Act's (MVRA) requirements imposing mandatory restitution have been considered. *Id.* at 2737-44. Defendants subject to the VWPA and MVRA may have modified their decision to plea or go to trial had they been aware of the statutes' requirements at the time of their case.

¹⁸¹ *See, e.g.*, VA. CODE ANN. § 46.2-811 (West 2011) (prohibiting coasting while driving downhill); *id.* § 46.2-816 (proposed legislation prohibiting a driver from following another car too closely); *id.* § 46.2-818 (prohibiting drivers from blocking ingress or egress from the premises of any service facility); *id.* § 46.2-821 (requiring drivers to yield the right-of-way before entering the highway); *id.* § 46.2-830 (requiring drivers to obey highway signs); *id.* § 46.2-838 (detailing the proper way to pass other moving motor vehicles).

¹⁸² *See, e.g., id.* § 46.2-395.

¹⁸³ DENNIS M. BROWN, ECON. RESEARCH SERV., PUBLIC TRANSPORTATION ON THE MOVE IN RURAL AMERICA (2008), available at <http://www.nal.usda.gov/ric/ricpubs/publictrans.htm> (noting that "only about a half of one percent of non[-]metro residents" use public transportation to get to work).

¹⁸⁴ *See, e.g.*, VA. CODE ANN. § 46.2-301 (West 2011) (detailing the consequences of driving on a suspended license).

¹⁸⁵ BORUCHOWITZ, BRINK & DIMINO, *supra* note 12, at 26.

quences attaching to a criminal conviction. By removing the impact of collateral consequences from this type of offense, legislatures could begin to limit the often debilitating impact of collateral sanctions.¹⁸⁶ Payment plan options for such infractions offer another viable alternative to the crushing debt that can attach to this offense.¹⁸⁷

Legislatures should also work to decriminalize minor drug possession charges. Convictions stemming from the possession of a controlled substance are often imposed for possessing any amount of the substance,¹⁸⁸ and these convictions carry heavy criminal penalties.¹⁸⁹ Significantly, women are more likely than men to face incarceration for non-violent drug offenses.¹⁹⁰ Because women reliably face drug offenses more frequently than men, and because other regulatory schemes exist to control drug use that does not victimize others¹⁹¹ without imposing criminal sanctions,¹⁹² existing punitive and prohibitory statutes should be modified.¹⁹³ Decriminalization of minor drug possession charges would not only reduce the burden on the court and prison systems, but would help to equalize the disparity in collateral consequences that male and female inmates face upon release.

Rather than focusing on decriminalization, the Second Chance for Ex-Offenders Act of 2011¹⁹⁴ is another potential solution to the excessive burden of collateral consequences. This bill has been introduced nearly every year since 2000 and would expunge first-time non-violent felony convictions.¹⁹⁵ Further, states that have experimented with the expungement of

¹⁸⁶ See *supra* Part III.A for a detailed explanation of the collateral consequences that attach to criminal convictions, and secondary impact of these consequences.

¹⁸⁷ Ashley Hooker, *Steps if Your Drivers License is Suspended for Failure to Pay Fines*, LEGAL AID OF W. VA., <http://www.lawv.net/node/1303> (last updated Jan. 18, 2011) (explaining the steps to take to sign up for West Virginia's payment plan); MARGY WALLER ET AL., BROOKINGS INST., DRIVER'S LICENSE SUSPENSION POLICIES 124-25 (2005), available at http://www.aecf.org/upload/publicationfiles/license_all_reports.pdf (explaining D.C.'s version of the payment plan).

¹⁸⁸ See, e.g., VA. CODE ANN. §§ 54.1-3446, 54.1-3448, 54.1-3450, 54.1-3452, 54.1-3454 (West 2011) (clarifying that any amount of a Schedule I, II, III, IV, or V substance is sufficient for a possession of a controlled substance charge under § 18.2-250).

¹⁸⁹ See, e.g., *id.* § 18.2-250 (imposing felony or misdemeanor convictions, depending on the type of drug, for possession of a controlled substance).

¹⁹⁰ BECK & HARRISON, *supra* note 145, at tbl. 17.

¹⁹¹ JOHN M. SCHEB & JOHN M. SCHEB II, CRIMINAL LAW & PROCEDURE 6 (6th ed. 2008).

¹⁹² Robert MacCoun et al., *Assessing Alternative Drug Control Regimes*, 15 J. OF POL'Y ANALYSIS & MGMT. 330, 334 (1996) (explaining less severe methods of addressing possession charges).

¹⁹³ Eric Blumenson & Eva Nilsen, *No Rational Basis: The Pragmatic Case for Marijuana Law Reform*, 17 VA. J. SOC. POL'Y & L. 43, 70-72, 74-75 (2009) (highlighting that rescheduling substances such as marijuana would serve to remove the sting of a more serious felony, as opposed to misdemeanor conviction, and that decriminalization in favor of fines could ameliorate the effect of collateral consequences altogether).

¹⁹⁴ H.R. 2065, 112th Cong. (2011).

¹⁹⁵ Thomas Kinney, *Congress Set to Dump "Second Chance for Ex-Offenders Act,"* THE-SLAMMER.ORG (Oct. 8, 2009), <http://www.the-slammer.org/carousel/congress-set-to-dump-second->

criminal records have been successful and can provide models that the federal government can draw on to reduce the impact of collateral consequences on ex-offenders.¹⁹⁶ Forgiving a first-time offender would lift the burden that collateral consequences place on such offenders. The reduction of this burden would allow for increased access to jobs and social services, such as student loans and food stamps, and it could break the cycle of recidivism into which so many offenders fall.

CONCLUSION

The costs of collateral consequences, such as bars to employment, loans, and federally subsidized housing, exceed their economic benefits. Further, the negative externalities resulting from collateral consequences damage society to such an extent that statutory provisions must be amended to internalize these effects. Female ex-offenders are particularly burdened by several of the collateral consequences that result from a criminal conviction. Solutions such as decriminalization, tying collateral consequences to ex post facto considerations, and legislation providing for expungement for first-time offenders can help to ameliorate the burden these consequences place on ex-offenders.

chance-for-ex-offenders-act (providing examples of non-violent felony charges, including conspiracy, dog fighting, obstruction of justice, and making false statements).

¹⁹⁶ Lahny R. Silva, *Clean Slate: Expanding Expungements and Pardons for Non-Violent Federal Offenders*, 79 U. CIN. L. REV. 155, 191, 195-96 (2010) (describing programs in Massachusetts, California, and Connecticut that have been successful in aiding ex-offender reintegration into various communities).

EMPOWERING LOCAL AND SUSTAINABLE FOOD:
DOES THE FOOD SAFETY MODERNIZATION ACT'S
TESTER–HAGAN AMENDMENT REMOVE ENOUGH BARRIERS?

*Peter Anderson**

INTRODUCTION

In recent years demand for local, sustainably produced food¹ has grown dramatically,² owing much to popular literature and films, as well as publicized outbreaks in foodborne illnesses linked to industrial agriculture.³ Consumers cite a broad range of motivating factors for switching to locally produced, non-industrial food: recent *E. coli* infections in spinach and peanut butter,⁴ salmonella in eggs,⁵ the inherent transparency of local food,⁶ fear of pasteurization,⁷ concerns about the energy consumption required to ship food across the country (or world),⁸ security of the food supply,⁹ and

* J.D. Candidate, 2013, George Mason University School of Law; B.A., 2004, James Madison University. The author would like to thank his family and friends for their support and helpful feedback.

¹ STEVE MARTINEZ ET AL., USDA, ECON. RESEARCH SERV., LOCAL FOOD SYSTEMS: CONCEPTS, IMPACTS, AND ISSUES i (2010), available at http://www.ers.usda.gov/media/122868/err97_1_1.pdf (“There is no consensus on a definition of ‘local’ or ‘local food systems’ in terms of the geographic distance between production and consumption. But defining ‘local’ based on marketing arrangements, such as farmers selling directly to consumers at regional farmers’ markets or to schools, is well recognized.”) (emphasis added).

² *Id.* at iii-iv; see also MICHAEL POLLAN, THE OMNIVORE’S DILEMMA: A NATURAL HISTORY OF FOUR MEALS 136 (2006) (“The word ‘organic’ has proved to be one of the most powerful words in the supermarket: Without any help from government, farmers and consumers working together . . . have built an \$11 billion industry that is now the fastest growing sector of the food economy.”).

³ Neil D. Hamilton, *Moving Toward Food Democracy: Better Food, New Farmers, and the Myth of Feeding the World*, 16 DRAKE J. AGRIC. L. 117, 123 (2011); Nathan M. Trexler, Comment, “Market” Regulation: Confronting Industrial Agriculture’s Food Safety Failures, 17 WIDENER L. REV. 311, 337 (2011).

⁴ A. Bryan Endres & Nicholas R. Johnson, *Integrating Stakeholder Roles in Food Production, Marketing, and Safety Systems: An Evolving Multi-Jurisdictional Approach*, 26 J. ENVTL. L. & LITIG. 29, 33 (2011).

⁵ *Id.*

⁶ JAMES T. O’REILLY, A CONSUMER’S GUIDE TO FOOD REGULATION & SAFETY 113 (2010) (“Food is produced more carefully by actual farmers—persons who either sell at a farmer’s market or whose identities are known to repeat customers in their local community.”).

⁷ The Weston A. Price Foundation, *Fresh, Unprocessed (Raw) Whole Milk: Safety, Health and Economic Issues*, REALMILK.COM (2009), <http://www.realmilk.com/rawmilkoverview.html>.

⁸ O’REILLY, *supra* note 6, at 113 (“Less energy is burned to move the crop from a distant area to the local consumer.”); POLLAN, *supra* note 2, at 133 (“I don’t believe it’s sustainable—or “organic” if you will—to FedEx meat all around the country.”). Americans purchasing produce grown in Argentina

new government initiatives that promote local purchasing.¹⁰ However, the interest in protecting consumers from foodborne pathogens through strict government regulation often conflicts with the interest in promoting local, non-industrial farming and processing.¹¹

On one hand, consumers expect the federal government to take any necessary steps to prevent dangerous or deadly food from entering the market,¹² with constituent outrage sometimes fueling new legislation or regulations.¹³ On the other hand, consumers sometimes look to alternatives like locally produced food in order to protect themselves.¹⁴ With people actively seeking out these alternatives, it has become apparent that uniform regulation of the nation's food supply may not be the best approach to food safety, considering the vastly different scales, designs, and goals of industrial agriculture compared to local independent farming.¹⁵ Members of Congress have acknowledged the difficulties that uniform regulation presents considering the competing goals of large-scale food safety and small business promotion.¹⁶ Indeed, local food producers are not able to

involves several "ethical implications . . . [t]here's the expense, there's the prodigious amounts of energy involved, the defiance of seasonality," etc. *Id.*

⁹ POLLAN, *supra* note 2, at 261 ("The important thing is that there be multiple food chains, so that when any one of them fails—when the oil runs out, when mad cow or other foodborne diseases become epidemic . . . we'll still have a way to feed ourselves."); see also O'REILLY, *supra* note 6, at 113 ("Retaining the capacity to sustain ourselves despite the globalization of commerce has a long-term cultural attractiveness.").

¹⁰ Hamilton, *supra* note 3, at 120; Trexler, *supra* note 3, at 341-42.

¹¹ See 156 CONG. REC. S8266-67 (daily ed. Nov. 30, 2010) (statement of Sen. Thomas Harkin) ("[O]ne of the most difficult issues I have had to face as manager of S. 510 is the balance between small growers and processors and larger producers and food companies.").

¹² Findings by the Bureau of Chemistry that certain food additives, such as formaldehyde, boric acid, salicylates, sulfites, and benzoates are dangerous to human health, along with publication of Upton Sinclair's famous skewering of the meat processing industry, *The Jungle*, inspired the Pure Food and Drug Act of 1906, the first federal law regulating adulterated food and drugs. PATRICIA A. CURTIS & WENDY DUNLAP, GUIDE TO FOOD LAWS AND REGULATIONS 28-31 (2005). The Act's replacement, the Federal Food, Drug, and Cosmetic Act of 1938, was also partly inspired by public outrage over the elixir sulfanilamide tragedy, in which over 100 people died from administration of a new drug that was not subject to premarket testing. *Id.* at 30-31.

¹³ Endres & Johnson, *supra* note 4, at 33-37; Denis Stearns, *Preempting Food Safety: An Examination of USDA Rulemaking and Its E. coli 0157:H7 Policy in Light of Estate of Kriefall ex rel. Kriefall v. Excel Corporation*, 1 J. FOOD L. & POL'Y 375 (2005); Eileen S. Pape, Comment, *A Flawed Inspection System: Improvements to Current USDA Inspection Practices Needed to Ensure Safer Beef Products*, 48 HOUS. L. REV. 421, 438-46 (2011).

¹⁴ Trexler, *supra* note 3, at 337.

¹⁵ David A. Taylor, *Does One Size Fit All? Small Farms and U.S. Meat Regulations*, 116 ENVTL. HEALTH PERSP. A529, A529 (2008).

¹⁶ 156 CONG. REC. S8266-67 (daily ed. Nov. 30, 2010) (statement of Sen. Thomas Harkin).

absorb the costs of new government regulations as readily as the industrial systems for which most regulations are designed.¹⁷

How can local food survive as an alternative to industrial agriculture if regulation designed to increase safety in industrial food production raises costs to the point that small farmers and processors can no longer compete?¹⁸ When Congress passed the FDA Food Safety Modernization Act (FSMA) in 2010, various advocates for the local food movement pushed for amendments exempting non-industrial farmers from several of the new safety measures imposed by the bill.¹⁹ Due in part to these organizations' lobbying,²⁰ Congress added the Tester–Hagan Amendment to the bill (H.R. 2751 in the House and S. 510 in the Senate) before it passed.²¹ The Tester–Hagan Amendment exempts farmers who gross less than \$500,000 annually and who sell at least 50% of their products directly to qualifying consumers—individuals, retailers located within 275 miles of the farm, or retailers located in the same state.²² However, some advocacy groups doubt that these exemptions go far enough to protect small farmers' businesses.²³ Other commentators point to potential manipulation of the FSMA by the FDA to allow other costly regulations that were already abandoned by the U.S. Department of Agriculture (USDA).²⁴ By contrast, some congressmen

¹⁷ POLLAN, *supra* note 2, at 249; Trexler, *supra* note 3, at 339-40; Sarah Breselor, *Will the USDA Doom Locally Produced Meat?*, SALON (Apr. 26, 2010, 11:27 A.M.), http://www.salon.com/food/feature/2010/04/26/usda_testing_end_local_meat.

¹⁸ Taylor, *supra* note 15, at A529, A531.

¹⁹ Helen Dombalis, *Tester – Now More Than Ever*, NAT'L SUSTAINABLE AGRIC. COAL., (Aug. 23, 2011), <http://sustainableagriculture.net/blog/tester-now-more-than-ever/> (“NSAC supported the Tester-Hagan Amendment and advocated for its inclusion in the final legislation, for the very reason that it allows smaller farms that sell products locally to play to their natural strengths in terms of food safety.”).

²⁰ *Update: VICTORY on S.510 Food Safety Bill*, VA. INDEP. CONSUMERS & FARMERS ASS'N (Nov. 21, 2010), <http://www.vicfa.org/news.html>; *Congress Passes Food Safety Bill with the Tester-Hagan Amendment*, FARM & RANCH FREEDOM ALLIANCE (Dec. 21, 2010), http://farmandranchfreedom.org/food_safety_bills_09; *Food Safety Action Alert*, NAT'L SUSTAINABLE AGRIC. COAL. (Nov. 10, 2010), <http://sustainableagriculture.net/blog/food-safety-action-alert-2/>.

²¹ 156 CONG. REC. S8264-65 (daily ed. Nov. 30, 2010) (statement of Sen. Sheldon Whitehouse) (“I am very pleased that all of this is accomplished while protecting small farmers and producers . . . I thank Senator Tester for his work on a compromise to protect farmers like those in Rhode Island, and throughout the Nation, who believe in the value of locally grown food.”).

²² *Congress Passes Food Safety Bill with the Tester-Hagan Amendment*, FARM & RANCH FREEDOM ALLIANCE (Dec. 21, 2010), http://farmandranchfreedom.org/food_safety_bills_09.

²³ Judith McGeary, *FDA Acts on Food Safety Bill*, FARM-TO-CONSUMER LEGAL DEF. FUND (May 20, 2011), <http://www.farmtoconsumer.org/fda-acts-on-food-safety-bill.htm>.

²⁴ David Gumpert, *The Hidden Agenda of the Tester-Hagan Amendment: Hint, It Has to Do with NAIS (Remember That?)*, THE COMPLETE PATIENT (Dec. 2, 2010, 12:53 P.M.), <http://www.thecompletepatient.com/journal/2010/12/2/the-hidden-agenda-of-the-tester-hagan-amendment-hint-it-has.html>.

expressed concern that including the Tester–Hagan Amendment would unnecessarily weaken the FSMA’s improvements to food safety.²⁵

This comment analyzes the potential problems posed by the FSMA and explores whether the Tester–Hagan Amendment’s exemptions for small producers adequately address their economic concerns. Part II summarizes the recent growth of the local food movement and its relevance to the consumer. It then briefly traces the history and evolution of food safety regulation in the U.S., including the FSMA and the Tester–Hagan Amendment. Part III describes the relief the Tester–Hagan Amendment provides to the local food movement and argues that despite its exemptions, the FSMA remains flawed because the revenue ceiling required for exemption discourages expansion of existing small farms and creates a barrier to entry for potential new farmers. Finally, this comment recommends that Congress amend the FSMA further to correct incentives while continuing to promote the overarching food safety goal.

I. BACKGROUND: THE LOCAL FOOD MOVEMENT AND FOOD SAFETY LAW

A. *The Growing Popularity and Rationale for Local Food*

Over the last twenty years, demand for locally produced food has risen steadily,²⁶ even though locally produced food remains a relatively small portion of the U.S. food economy as a whole.²⁷ Despite higher market prices for locally produced items compared to industrially produced items, the typical outlets for local food, including farmers’ markets, direct sales to restaurants, and community supported agriculture (CSA) programs,²⁸ have

²⁵ 156 CONG. REC. H8887 (daily ed. Dec. 21, 2010) (statement of Rep. Henry Waxman) (“Some on the other side of the aisle, Republicans, are saying we should reject the whole bill because of the Tester amendment, which exempts small farmer-producers and facilities. We didn’t have that in our bill, and I would have preferred that the Senate had not adopted that provision. But I don’t think it is a reason to vote against this whole bill.”).

²⁶ MARTINEZ ET AL., *supra* note 1, at iii (2010) (“Direct-to-consumer marketing amounted to \$1.2 billion in current dollar sales in 2007, according to the 2007 Census of Agriculture, compared with \$551 million in 1997.”); Taylor, *supra* note 15, at A529, A529.

²⁷ MARTINEZ ET AL., *supra* note 1, at iii (2010) (“Direct-to-consumer sales accounted for 0.4 percent of total agricultural sales in 2007, up from 0.3 percent in 1997.”).

²⁸ Trexler, *supra* note 3, at 338. CSA programs involve the purchase of shares of a harvest for a particular growing season in advance—with produce either delivered directly to the shareholders or arranged for pick-up—often on a weekly basis during the particular season. Marne Coit, *Jumping on the Next Bandwagon: An Overview of the Policy and Legal Aspects of the Local Food Movement*, 4 J. FOOD L. & POL’Y 45, 59-60 (2008). CSAs are one tool making local food accessible to urban communities. See POLLAN, *supra* note 2, at 245.

grown significantly.²⁹ Since 1994, the number of active farmers' markets has more than doubled.³⁰ Reasons for this shift to local food as an alternative to cheaper and more readily available industrial food include perceived safety and health benefits, as well as sustainability concerns.³¹ Consumers also report a preference for local produce due to perceived freshness, a desire to support local businesses, and a desire to know the source of their food.³² Also affecting consumers' decisions are the infamous *E. coli* contaminations in hamburgers and bagged spinach,³³ as well as Salmonella outbreaks in alfalfa, eggs, and peanut butter that caused numerous deaths and illnesses.³⁴ Furthermore, a Listeria outbreak from contaminated cantaloupes killed at least twenty-eight people in the United States in 2011.³⁵

Many perceive locally produced food as a safer alternative because of increased transparency,³⁶ reduced time in storage and transport, and reduced or zero use of chemical pesticides and antibiotics.³⁷ Moreover, the conventional logic dictates that food traveling shorter distances³⁸ is more environmentally responsible, or "sustainable," because less fossil fuel is required to move the products to market.³⁹ Reduced travel and processing also allows fewer opportunities for contamination.⁴⁰ However, these incentives to purchase local food are tempered with costs; evidence suggests that more

²⁹ MARTINEZ ET AL., *supra* note 1, at iii ("The number of farmers' markets rose to 5,274 in 2009, up from 2,756 in 1998 and 1,755 in 1994 . . . In 2005, there were 1,144 community-supported agriculture organizations (CSAs) in operation, up from 400 in 2001 and 2 in 1986 . . ."); Taylor, *supra* note 15, at A529.

³⁰ Taylor, *supra* note 15, at A529.

³¹ *Id.*

³² MARTINEZ ET AL., *supra* note 1, at 29.

³³ Endres & Johnson, *supra* note 4, at 37 (Several hundred people suffered illnesses from bagged spinach tainted with *E. coli*); Taylor, *supra* note 15, at A529 (*E. coli* in hamburgers killed four people in a 1992 Jack-in-the-Box outbreak.).

³⁴ Endres & Johnson, *supra* note 4, at 33-34.

³⁵ *Death Toll From Listeria Outbreak Rises to 28*, REUTERS (Oct. 26, 2011, 11:13 AM), <http://www.reuters.com/article/2011/10/26/us-death-toll-listeria-idUSTRE79P51R20111026>.

³⁶ The transparency argument presumes that "[f]ood is produced more carefully by actual farmers—persons who either sell at a farmer's market or whose identities are known to repeat customers in the local community." O'REILLY, *supra* note 6, at 113.

³⁷ Taylor, *supra* note 15, at A529; Trexler, *supra* note 3, at 338.

³⁸ There is no legal consensus on what distance qualifies food as "local." MARTINEZ ET AL., *supra* note 1, at 3. The New Oxford American Dictionary defines a "locavore" as someone who consumes food produced within a 100-mile radius of her residence, whereas the 2008 Farm Bill defines food qualifying for assistance under the USDA Value-Added Agricultural Development program as "locally produced" if it travels less than 400 miles from producer to consumer or is consumed within the same state. *Id.* The FSMA splits the difference, with a "qualifying end user" being an individual who consumes the food, or a restaurant located within the same state as the producer or within 275 miles of the producer. 21 U.S.C. § 350g(l)(4) (Supp. IV 2010).

³⁹ Taylor, *supra* note 15, at A529; *see also* POLLAN, *supra* note 2, at 175 ("I don't believe it's sustainable—or 'organic' if you will—to FedEx meat all around the country.").

⁴⁰ Trexler, *supra* note 3, at 338.

people would be inclined to purchase locally produced food if it were more accessible, more convenient, and lower in price.⁴¹

Increased public awareness of local food options may be attributed to the work of advocacy groups, like the National Sustainable Agriculture Coalition; popular books and films, such as *Food, Inc.* and books by Michael Pollan; and programs initiated by the U.S. government itself, like the USDA's "Know Your Farmer Know Your Food" campaign.⁴² Corporations like Wal-Mart have also acknowledged the importance of sustainability through the purchase and retail sale of locally produced foods.⁴³

A central figure in Michael Pollan's bestselling book, *The Omnivore's Dilemma: A Natural History of Four Meals*,⁴⁴ Joel Salatin, is a Virginia farmer who primarily raises cattle, pigs, chickens, and eggs. Salatin continually moves his livestock to different grass pastures in a sequence that allows him to raise them without using hormones or antibiotics.⁴⁵ He sells all of his products on-site to visiting customers, with the exception of deliveries made to nearby restaurants.⁴⁶ Salatin's system demonstrates the kind of success an independent farmer running a small operation can have, and his media exposure—Pollan's book, significant screen time in *Food, Inc.*, various interviews, and his own books and articles—has made him both a minor folk hero and helped drive the local food movement.⁴⁷

There is anecdotal evidence not only of increased interest in consuming local food, but also in entering local farming as an occupation.⁴⁸ Diverse groups of people are attending conferences and pursuing internships in farming—often people who were not raised on farms and possess non-traditional ideas about agriculture.⁴⁹ Joel Salatin himself provides summer internships and apprenticeships for those who want to learn his sustainable

⁴¹ MARTINEZ ET AL., *supra* note 1, at 30 ("Surveys suggest that reasons for not shopping at a farmers' market include: absence of availability in the patron's vicinity; lack of knowledge about market existence; inconvenience (too far to drive); food of comparable quality at more convenient locations; and prices being too high (possibly due to timing of survey—beginning of the season).").

⁴² Hamilton, *supra* note 3, at 118-21.

⁴³ *Id.* at 122-23 (citing WAL-MART STORES, INC., WAL-MART GLOBAL SUSTAINABILITY REPORT: 2010 PROGRESS UPDATE 5 (2010), *available at* <http://cdn.walmartstores.com/sites/sustainabilityreport/2010/WMT2010GlobalSustainabilityReport.pdf>).

⁴⁴ POLLAN, *supra* note 2, at 123-261.

⁴⁵ *Id.* at 125-33.

⁴⁶ *Id.* at 240.

⁴⁷ Andrea Gabor, *Inside Polyface Farm, Mecca of Sustainable Agriculture*, THE ATLANTIC (July 25, 2011, 12:31 PM), <http://www.theatlantic.com/life/archive/2011/07/inside-polyface-farm-mecca-of-sustainable-agriculture/242493/>.

⁴⁸ Hamilton, *supra* note 3, at 123; Dan Charles, *Newbie Farmers Find That Dirt Isn't Cheap*, THE SALT: NPR'S FOOD BLOG (Nov. 15, 2011, 10:27 AM), <http://www.npr.org/blogs/thesalt/2011/11/14/142305869/newbie-farmers-find-that-dirt-isnt-cheap?sc=emaf>.

⁴⁹ Hamilton, *supra* note 3, at 129.

style of farming,⁵⁰ as well as a “field day” where the public may tour his farm and learn about his particular style of farming.⁵¹

Despite this surge in smaller-scale farming interest, new farmers face significant barriers to entry.⁵² Besides raising capital and finding suitable land,⁵³ the time and labor the small farmer must devote to direct marketing can inhibit growth of the farm operation.⁵⁴ The USDA’s Economic Research Service (ERS) reports lack of affordable distribution through local supply chain infrastructures, recordkeeping requirements, education and training, and regulatory uncertainties as additional barriers to entry and expansion for local farmers.⁵⁵

The federal government has addressed some of these barriers directly.⁵⁶ In September 2009, Secretary of Agriculture Tom Vilsack introduced the USDA’s “Know Your Farmer Know Your Food” campaign.⁵⁷ He advocated for a shift back to the small farmer system of pre-industrial agriculture when he addressed the Senate Committee on Agriculture, Nutrition and Forestry in June 2010:

Let me suggest one idea that this committee might consider. Why not set as a goal for the 2012 Farm Bill the ability to add at least 100,000 additional farmers in the area of the small farming and commercial operations? Why not establish local advisory councils in communities across the country [to] identify, recruit, encourage and assist young people to consider a life in farming?⁵⁸

This is not the first time the federal government has attempted to connect consumers to the farmers themselves instead of fostering a chain of processing and distribution networks separating the animal or plant being eaten from the person eating it. In 1976, Congress passed the Farmer-to-Consumer Direct Marketing Act to help the development of local food deli-

⁵⁰ POLYFACE: “THE FARM OF MANY FACES,” <http://www.polyfacefarms.com/apprenticeship/> (last visited Nov. 18, 2011).

⁵¹ Gabor, *supra* note 47.

⁵² Charles, *supra* note 48 (Respondents to a National Young Farmers’ Coalition survey stated that their biggest challenges were “lack of capital and land access.”).

⁵³ *Id.*

⁵⁴ MARTINEZ ET AL., *supra* note 1, at 23 (“In other words, the incentive of smaller farmers to expand and become more efficient is diminished as more time is spent off-farm performing additional entrepreneurial activities such as marketing at farmers’ markets.”).

⁵⁵ *Id.* at 23-27.

⁵⁶ For example, the USDA provides Hazard Analysis and Critical Control Point (HACCP) training and education materials for small meat producers. Taylor, *supra* note 15, at A531.

⁵⁷ USDA, News Release, *USDA Launches ‘Know Your Farmer, Know Your Food’ Initiative to Connect Consumers with Local Producers to Create New Economic Opportunities for Communities* (Sept. 15, 2009), <http://www.usda.gov/wps/portal/usda/usdahome?contentidonly=true&contentid=2009/09/0440.xml>.

⁵⁸ Hamilton, *supra* note 3, at 127-28.

very systems like farmers' markets and CSAs.⁵⁹ This law requires the USDA to fund direct-to-consumer conferences and assist state legislatures in creating their own direct-to-consumer marketing programs.⁶⁰ While Secretary Vilsack was not the first federal department head to support non-industrial farming, he has reaffirmed the government's commitment, at least in principle, to the local food movement.⁶¹

The primary goal of federal food regulation is safety.⁶² Supporting the local food movement serves this goal because anecdotal evidence demonstrates that outbreaks of foodborne illnesses almost always derive from industrial agriculture rather than local agriculture.⁶³

B. *A Brief History of U.S. Government Regulation of Food Safety*

When Abraham Lincoln created the U.S. Department of Agriculture in 1862, 48% of Americans lived on farms.⁶⁴ By 2000, only 1 % of Americans lived on farms.⁶⁵ In the interim, the U.S. food economy shifted towards specialization and globalization driven by new technologies.⁶⁶ Only recently has the food economy shown some signs of "relocalization."⁶⁷

Federal regulation of food safety began with the 1906 Pure Food & Drug Act, which charged the USDA's Bureau of Chemistry (later renamed the Food and Drug Administration (FDA)) with enforcement.⁶⁸ That same year, the Meat Inspection Act gave the USDA jurisdiction over prescribing and enforcing sanitation standards for slaughterhouses and processing plants.⁶⁹ In 1938, the Federal Food, Drug, and Cosmetic Act replaced the Pure Food & Drug Act, and two years later the FDA transferred out of the

⁵⁹ Trexler, *supra* note 3, at 340-41.

⁶⁰ *Id.* at 341.

⁶¹ Hamilton, *supra* note 3, at 128.

⁶² Endres & Johnson, *supra* note 4, at 36 ("[O]ur population is dependent upon mass producers for its food and drink . . . [that are no longer] natural or simple products but complex ones whose composition and qualities are often secret. Such a dependent society must exact a greater care than in more simple days and must require from manufacturers or producers increased integrity and caution as the only protection of its safety and well-being." (quoting *Dalehite v. United States*, 346 U.S. 15, 51-52 (1953) (Jackson, J., dissenting))).

⁶³ Trexler, *supra* note 3, at 339 ("In 16 years, I've never had an outbreak linked to a farmers market." (quoting attorney William Marler, who represents victims of foodborne illnesses) (citing Kristin Choo, *Hungry for Change*, A.B.A. J., Sept. 2009, at 61)).

⁶⁴ CURTIS & DUNLAP, *supra* note 12, at 33 (Lincoln referred to the USDA as the "people's department.").

⁶⁵ MARTINEZ ET AL., *supra* note 1, at 1.

⁶⁶ *Id.*

⁶⁷ *Id.*

⁶⁸ Endres & Johnson, *supra* note 4, at 41.

⁶⁹ *Id.*

USDA⁷⁰ to the Federal Security Agency.⁷¹ Since 1979, the FDA has been under the purview of the Department of Health and Human Services.⁷²

One author asserts that because enforcement responsibilities are often delegated—and statutes are often written—in response to specific safety calamities, the resulting regulatory regime “is somewhat a haphazard patchwork.”⁷³ The USDA and the FDA currently share primary responsibility for ensuring the safety of the nation’s food supply,⁷⁴ but a number of other federal agencies are also involved.⁷⁵ This department overlap can lead to confusion and inefficiency.⁷⁶

Under the 1957 Poultry Products Inspection Act and the 1970 Egg Products Inspection Act (including 1991 amendments), the USDA currently has responsibility for the safety of meat, poultry, and eggs that are broken for processing into other products, while the FDA has jurisdiction over “shell eggs,” which have not been broken for processing, along with virtually every other kind of food.⁷⁷ Under the current regime, the USDA inspects processing plants in its jurisdiction at least four times per year.⁷⁸ The USDA also mandated Hazard Analysis and Critical Control Point (HACCP) plans for meat and poultry processors in 1998, with full implementation required by 2000.⁷⁹ HACCP plans aim to identify the riskiest points along the production process and create breaks in production for cleaning.⁸⁰ The plans also set precise rules for methods of cooking in order to minimize the risks of food contamination or, at a minimum, isolate any problems that are found.⁸¹ HACCP plans focus on prevention, rather than “end product testing,” to achieve greater safety than inspection-only systems.⁸² These plans

⁷⁰ *Id.* at 41-42.

⁷¹ CURTIS & DUNLAP, *supra* note 12, at 34.

⁷² *Id.*

⁷³ NEIL D. FORTIN, *FOOD REGULATION: LAW, SCIENCE, POLICY, AND PRACTICE* 27 (2009) (“Just as the statutes were written to address specific problems at particular points in history, the delegation of food regulation was developed to address specific concerns. The delegation therefore represents an evolution rather than an organization by design.”); *see also* CURTIS & DUNLAP, *supra* note 12, at 28-34.

⁷⁴ Endres & Johnson, *supra* note 4, at 41-42.

⁷⁵ FORTIN, *supra* note 73, at 27.

⁷⁶ Endres & Johnson, *supra* note 4, at 42 (“[T]he FDA has sole jurisdiction over the manufacturing of a cheese pizza, while the USDA has sole authority over a meat pizza by virtue of the meat content.”). It should also be noted that while the USDA and FDA administer national food safety policy, in practice it is state and local health and food safety agencies that work “closest to the point of food delivery to the consumer It has been estimated that 80 percent of food safety inspections are done by these local agencies.” O’REILLY, *supra* note 6, at 9.

⁷⁷ Endres & Johnson, *supra* note 4, at 37, 41-43.

⁷⁸ *Id.*

⁷⁹ FORTIN, *supra* note 73, at 244.

⁸⁰ Taylor, *supra* note 15, at A530.

⁸¹ *Id.*

⁸² FORTIN, *supra* note 73, at 242 (“HACCP prevents foodborne illness by applying science to identify risks in a method of food handling or processing. It controls those risks through preventative

are required for every food processor that comes under the USDA's jurisdiction.⁸³

One problem developing for the last thirty years, however, is the consolidation of meat processors and therefore USDA-approved inspection sites, which are administered through the Food Safety and Inspection Service (FSIS) branch of the USDA.⁸⁴ Since 1981, the number of slaughterhouses and FSIS inspectors have each declined by approximately 10%.⁸⁵ During the same period, however, meat and poultry production doubled.⁸⁶ This suggests that meat processing facilities are now larger, but there are fewer FSIS inspectors. As a result, "the number of FSIS employees per billion pounds of meat and poultry inspected and approved has declined by more than half since 1981."⁸⁷

A collateral effect of this trend has been a less safe meat supply, as demonstrated by increased frequency of *E. coli* outbreaks.⁸⁸ Another consequence of the centralization of meat inspection is that small cattle and poultry farmers must ship their meat greater distances to the nearest USDA-approved inspection facility, and therefore they must absorb the higher costs associated with the travel.⁸⁹ This threatens to take the "local" out of locally raised meat.⁹⁰ A surprising result is that meat purchased from a local farmer may actually have traveled more miles to and from the inspector than a more distant farmer's meat travels for distribution to consumers, simply by virtue of the latter's proximity to a USDA inspection facility.⁹¹ By contrast, shell eggs and other non-meat food products did not share this problem prior to the FSMA's passage because the FDA rarely visited shell egg facilities for inspection.⁹² Some commentators praise the FSMA for increasing the scope of the FDA's inspection power because unlike the

controls. Finally, HACCP is a complete system that includes corrective actions, record keeping, and verification, which increase the effectiveness and efficiency of both HACCP and conventional sanitation methods.").

⁸³ 9 C.F.R. § 417.2 (2011); *see also* Taylor, *supra* note 15, at A530.

⁸⁴ Taylor, *supra* note 15, at A530.

⁸⁵ *Id.*

⁸⁶ *Id.*

⁸⁷ *Id.* (quoting Michael Pollan).

⁸⁸ USDA, RECALL RELEASE, UPDATED: NEW JERSEY FIRM EXPANDS RECALL OF GROUND BEEF PRODUCTS DUE TO POSSIBLE E. COLI O157:H7 CONTAMINATION (Oct. 6, 2007), http://www.fsis.usda.gov/PDF/Recall_040_2007_Exp_Update.pdf. By contrast, local food is offered as a solution to the safety problems posed by centralized meat processing. MARTINEZ ET AL., *supra* note 1, at 42 ("It has been suggested that local food systems could reduce food safety risks by decentralizing production.").

⁸⁹ Taylor, *supra* note 15, at A530.

⁹⁰ *Id.* at A530-A531.

⁹¹ *Id.*

⁹² Endres & Johnson, *supra* note 4, at 43-44.

USDA, which has multiple missions, including marketing, the FDA is primarily a safety agency.⁹³

C. *The FDA Food Safety Modernization Act of 2010 and the Tester–Hagan Amendment*

Safety was the primary goal driving the FSMA’s passage.⁹⁴ The legislation was the first major update to the Food, Drug, and Cosmetic Act’s food regulation in seventy years.⁹⁵ The FSMA strengthens and consolidates the FDA’s ability to implement food safety rules.⁹⁶ It creates new regulatory power in the FDA to mandate safety measures for fruit and vegetable farming, install Hazard Analysis and Critical Control Points (HACCPs) for all non-meat food processing facilities,⁹⁷ oversee transportation of food, and inspect and copy all operational records kept at a facility when the agency investigates any food emergency which it “reasonably believes will cause serious adverse health consequences . . . to humans or animals.”⁹⁸ Further, the FSMA provides the FDA with the authority to issue a food recall whenever there is “reasonable probability” that a food is adulterated or misbranded and when exposure “will cause serious adverse health consequences or death” to humans or animals.⁹⁹

After extensive lobbying, advocates of the local food movement persuaded Congress to exempt non-industrial farmers and processors from some of the new regulations imposed by the FSMA.¹⁰⁰ The Tester–Hagan

⁹³ *Id.* at 44-45 (pointing out that the USDA’s Agricultural Marketing Service (AMS) does not even have safety as part of its mission statement, hence broader oversight powers for the FDA, which is a safety agency under HHS, correctly delegates food safety regulation).

⁹⁴ 156 CONG. REC. H8821 (daily ed. Dec. 21, 2010) (statement of Rep. Jim McGovern) (“Each year, 76 million Americans are sickened from consuming contaminated food, more than 300,000 people are hospitalized, and 5,000 die. In just the last few years, there has been a string of foodborne illness outbreaks in foods consumed by millions of Americans each day—from contaminated spinach to peanut butter to cookie dough. This bill puts a new focus on preventing food contamination before it occurs—putting new responsibilities on food producers and requiring them to develop a food safety plan and ensure the plan is working.”).

⁹⁵ 156 CONG. REC. H8885 (daily ed. Dec. 21, 2010) (statement of Rep. Frank Pallone) (“The modernization of our food safety system is desperately needed. The current food regulatory regime was established in 1938 and hasn’t been overhauled in 70 years.”); Amy Tsui, *Senate Passes Bill on Food Safety, Clearing Way for Consideration in House*, BNA INT’L TRADE REP. (Dec. 2, 2010), http://news.bna.com/itln/ITLNWB/split_display.adp?fedfid=18680981&vname=itrmotalissues&wsn=496505000&searchid=15930224&doctypeid=1&type=date&mode=doc&split=0&scm=ITLNWB&pg=0.

⁹⁶ Endres & Johnson, *supra* note 4, at 103-104.

⁹⁷ *Id.* at 106-107 (“[T]he FSMA did not significantly redistribute authority from USDA to FDA; it merely extended FDA’s jurisdiction . . .”).

⁹⁸ *Id.* at 103-104.

⁹⁹ 21 U.S.C. § 350(a) (2011).

¹⁰⁰ *Congress Passes Food Safety Bill with the Tester-Hagan Amendment*, FARM & RANCH FREEDOM ALLIANCE (Dec. 21, 2010), http://farmandranchfreedom.org/food_safety_bills_09.

Amendment exempts businesses that gross less than \$500,000 annually in food sales and sell more than half of their products directly to individual consumers or in-state retailers, or retailers within 275 miles of the producer, from: (1) previously existing registration mandates, (2) HACCP requirements, and (3) the FSMA's new "produce safety standards," which govern the entire farming process from planting to harvest.¹⁰¹ The size and revenue requirements that qualify producers for the exemption are as follows:

(1) IN GENERAL—A farm shall be exempt from the requirements under this section in a calendar year if—

(A) during the previous 3-year period, the average annual monetary value of the food sold by such farm directly to qualified end-users during such period exceeded the average annual monetary value of the food sold by such farm to all other buyers during such period; and

(B) the average annual monetary value of all food sold during such period was less than \$500,000, adjusted for inflation.¹⁰²

Of particular relevance for supporters of the local food movement is Congress's definition of a "qualified end-user" and "consumer." By "qualified end-user," Congress means:

(i) the consumer of the food; or (ii) a restaurant or retail food establishment (as those terms are defined by the Secretary for purposes of section 350d of this title) that is located-- (I) in the same State as the farm that produced the food; or (II) not more than 275 miles from such farm.¹⁰³

Congress clarified that here, "the term 'consumer' does not include a business."¹⁰⁴

The inclusion of a small farm and producer exemption did not garner universal approval in Congress.¹⁰⁵ It is important to remember, however, that the FSMA, while granting broad new powers of inspection and rule-making, along with the exemptions aimed at small local producers, empowers only the FDA, which lacks the authority to inspect meat, poultry, and

¹⁰¹ Judith McGeary, *Analysis of the Tester-Hagan Amendment*, FARM & RANCH FREEDOM ALLIANCE (Dec. 2, 2010), <http://farmandranchfreedom.org/Tester-Hagan-explanation>; 21 U.S.C. § 350h(f)(4)(A) (2006).

¹⁰² 21 U.S.C. § 350h(f)(1) (2006).

¹⁰³ § 350h(f)(4)(A).

¹⁰⁴ § 350h(f)(4)(B).

¹⁰⁵ See generally 156 CONG. REC. S8266-67 (daily ed. Nov. 30, 2010) (Senate debate preceding passage of the Senate's version of the bill, S. 510); 156 CONG. REC. H8884-90 (daily ed. Dec. 21, 2010) (House of Representatives debate preceding passage of H.R. 2751 with the Senate amendments).

processed egg products.¹⁰⁶ These products are still subject to USDA inspection and HACCP plans.¹⁰⁷

II. ECONOMIC EFFECTS: MATCHING THE STATUTE TO STATED POLICY GOALS

A. *How the FSMA Affects Local Food*

The Tester–Hagan Amendment exempted qualifying food producers from some of the more costly requirements of the FSMA.¹⁰⁸ Nearly as important, the Tester–Hagan Amendment mandates that the FDA study the rates of food-contaminating bacteria in products from small producers to determine whether the size of the farm or processing facility tends to affect the cleanliness of the products.¹⁰⁹ Few studies of this kind for meat and poultry have been conducted,¹¹⁰ but there is anecdotal evidence supporting the hypothesis that non-industrial meat is safer than industrial meat.¹¹¹ For example, Virginia’s Joel Salatin had his poultry independently tested and compared with industrial poultry when he was challenging a state inspector’s finding that his open-air poultry slaughter area was unsanitary.¹¹² The tests revealed that Salatin’s chickens contained an average of ten times fewer bacteria than the USDA-approved supermarket chickens.¹¹³

If the findings of the FSMA-mandated study support what many already suspect—namely, that small operations that sell locally are inherently safer¹¹⁴—it could vindicate HACCP exemptions for qualifying producers and perhaps even inspire similar exemptions for small meat processors under the USDA’s rules. All meat producers, local or industrial, must use

¹⁰⁶ Endres & Johnson, *supra* note 4, at 45.

¹⁰⁷ *Id.*

¹⁰⁸ 21 U.S.C. §§ 350g(l), 350h(f) (2011).

¹⁰⁹ *Id.* § 350g(l)(5)(iv); Judith McGeary, *Analysis of the Tester-Hagan Amendment*, FARM & RANCH FREEDOM ALLIANCE (Dec. 2, 2010), <http://farmandranchfreedom.org/Tester-Hagan-explanation> (“Directs FDA to conduct a study in the next 18 months to look at the incidence of foodborne illness in relation to food producers’ scale and type of operation. Directs FDA to use this study to define ‘very small businesses’ that will also be exempt from the HACCP-type requirements.”).

¹¹⁰ Taylor, *supra* note 15, at A531 (“There are few studies of the comparative health risks posed by small versus large processing facilities.”).

¹¹¹ *Id.* (Michael Pollan “argued that food safety problems from small players are ‘less catastrophic and easier to manage because local food is inherently more traceable and accountable.’”).

¹¹² *Id.* at A530-31.

¹¹³ *Id.* at A531.

¹¹⁴ Michael Pollan & Eric Schlosser, *A Stale Food Fight*, N.Y. TIMES (Nov. 28, 2010), <http://www.nytimes.com/2010/11/29/opinion/29schlosser.html?scp=1&sq=pollan%20schlosser%20a%20stale%20food%20fight&st=cse> (“The largest [foodborne illness] outbreaks are routinely caused by the largest processors, not by small producers selling their goods at farmers’ markets.”).

USDA processors, which employ the costly mandated HACCP rules.¹¹⁵ However, because the FSMA requires the FDA to collaborate with the USDA in studying the relationship between the size of the producer and the rate of foodborne illness,¹¹⁶ the USDA might find a basis for exempting its smaller processing facilities. When the House of Representatives passed the FSMA with the Senate amendments, including the Tester–Hagan Amendment, some members of the House advocated for changes in the USDA’s policy similar to the changes being passed for the FDA that day.¹¹⁷ Pending the results of the study, non-industrial meat processors might reap the benefits of Tester–Hagan-type exemptions in the future.¹¹⁸

In the meantime, all meat processing facilities are subject to HACCP.¹¹⁹ Salatin argues that while he must charge more than \$1 per pound more for his meats than industrial producers, several hidden costs exist in industrially produced meat, which are not reflected in the sale price to consumers.¹²⁰ These costs are not borne by the farmer, but by society at large, including the costs of “water pollution, of antibiotic resistance, of foodborne illnesses, of crop subsidies, [and] of subsidized oil and water.”¹²¹ If non-industrial farmers like Salatin were not required to subject their “clean” meats to standardized USDA inspection—which increases their costs, mainly due to the extra transportation involved—arguably they could price their meats at the market price for industrial meats.¹²²

In addition to mandating USDA collaboration on the study of the relationship between producer size and rate of foodborne illness, the FSMA encourages further collaboration between the FDA and the USDA for enforcement of the new produce safety standards to be written by the FDA.¹²³

¹¹⁵ See O’REILLY, *supra* note 6, at 10, 12.

¹¹⁶ 21 U.S.C. § 350g(l)(5) (2011) (“The Secretary, in consultation with the Secretary of Agriculture, shall conduct a study of the food processing sector regulated by the Secretary to determine . . . (iv) the incidence of foodborne illness originating from each size and type of operation and the type of food facilities for which no reported or known hazard exists.”).

¹¹⁷ See 156 CONG. REC. H8887 (daily ed. Dec. 21, 2010) (statement of Rep. Rosa DeLauro) (“While the FDA is charged with protecting a large majority of our food supply, the Food Safety and Inspection Service, FSIS at USDA, is responsible for ensuring the safety of meat and poultry products. After passing this bill today, we must begin to lay the foundation for science-based reform at FSIS as well.”).

¹¹⁸ *Id.*

¹¹⁹ 9 C.F.R. § 417.2 (2011); see also FORTIN, *supra* note 73, at 244 (“In 1998, the U.S. Department of Agriculture established HACCP for meat and poultry processing plants. . . . In 2000, FSIS completed implementation of its landmark rule on Pathogen Reduction and Hazard Analysis and Critical Control Point (HACCP) systems.”).

¹²⁰ POLLAN, *supra* note 2, at 243.

¹²¹ *Id.*

¹²² *Id.*

¹²³ 21 U.S.C. § 350h(d) (2011) (“The Secretary may coordinate with the Secretary of Agriculture and, as appropriate, shall contract and coordinate with the agency or department designated by the Governor of each State to perform activities to ensure compliance with this section.”).

There is an important economic difference between process standards and performance standards, yet the FSMA establishes both.¹²⁴ Process standards prescribe specific manufacturing or harvesting procedures and are considered less efficient than performance standards.¹²⁵ By contrast, performance standards set limits on acceptable levels of pathogens at a given point of production and are considered more efficient because producers, over time, can determine facility-specific paths to least cost compliance.¹²⁶ HACCP-type plans involve elements of both process and performance standards,¹²⁷ but the good news for non-industrial farmers is that the Tester–Hagan Amendment exempts them from both the HACCP-type regulations and the produce safety standards—assuming they qualify.¹²⁸ While collaboration between the FDA and the USDA may create greater flexibility of compliance for local meat producers in the future, the Tester–Hagan exemptions in the FSMA relieve qualifying farmers from the costly “Hazard Analysis and Risk Based Preventive Controls” and current standards for produce safety.¹²⁹

The Congressional Budget Office estimates that—because food producers themselves will be responsible for bearing the cost of the new regulations—the net cost to the U.S. government will be negligible, despite the new oversight responsibilities of the FDA.¹³⁰ Moreover, the FDA itself

¹²⁴ Compare 21 U.S.C. § 350g (requiring HACCP-type plans that will govern production methods), and *id.* § 350h (granting the FDA authority to prescribe specific production and harvesting procedures), with *id.* § 2201 (granting the FDA authority to establish foodborne contaminant performance standards).

¹²⁵ Laurian J. Unnevehr & Helen H. Jensen, *Industry Costs to Make Food Safe: Now and under a Risk-Based System*, in *TOWARD SAFER FOOD: PERSPECTIVES ON RISK AND PRIORITY SETTING* 105, 110–11 (Sandra A. Hoffman & Michael R. Taylor eds., 2005).

¹²⁶ *Id.*

¹²⁷ *Id.* at 115 (“[T]he nature of HACCP regulation is unclear—is it a performance standard or a process standard? [Some economists] describe the Pathogen Reduction Regulation in meat and poultry as a combination of performance and process standards.”).

¹²⁸ 21 U.S.C. §§ 350g(l), 350h(f) (2011). It is unclear whether FDA standards for produce safety issued under §350h will be shaped more as process standards or performance standards. See *id.* §350h(a)(1)(A) (“... the Secretary . . . shall publish a notice of proposed rulemaking to establish science-based minimum standards for the safe production and harvesting of fruits and vegetables . . . that are raw agricultural commodities for which the Secretary has determined that such standards minimize the risk of serious adverse health consequences or death.”). Rules for both the §350g preventive controls and produce safety under §350h were due in July 2012. See *Preventive Standards*, FDA, <http://www.fda.gov/Food/FoodSafety/FSMA/ucm256826.htm> (last updated Mar. 1, 2012). At the time of this writing, however, the FDA has not promulgated either of them. See *Implementation and Progress*, FDA, <http://www.fda.gov/Food/FoodSafety/FSMA/ucm250568.htm> (last updated Aug. 31, 2012). It should also be noted that there are no exemptions in the FSMA from the foodborne contaminant performance standards; they apply broadly to product classes and are not facility-specific. 21 U.S.C. § 2201 (2011).

¹²⁹ 21 U.S.C. §§ 350g(l), 350h(f) (2011).

¹³⁰ CONG. BUDGET OFFICE, CBO ESTIMATE OF THE STATUTORY PAY-AS-YOU-GO EFFECTS FOR H.R. 2751, FDA FOOD SAFETY MODERNIZATION ACT, AS PASSED BY THE SENATE ON DECEMBER 19,

acknowledges the importance of food producers bearing the costs of food safety, as opposed to the taxpayer.¹³¹ Penalties will only be assessed to those producers whose noncompliance results in recalls, facility re-inspections, or importer re-inspections, though producers will bear the costs of compliance.¹³²

While food producers not qualifying for exemption bear the costs of the FSMA, the American public will likely benefit a great deal from stricter regulation.¹³³ According to the Centers for Disease Control and Prevention, “[a]bout 48 million people (1 in 6 Americans) get sick, 128,000 are hospitalized, and 3,000 die each year from food-borne disease.”¹³⁴ One study has shown that the healthcare costs associated with these illnesses amount to roughly \$152 billion annually.¹³⁵ If producers come into compliance, not only will the public enjoy increased food safety, but those producers in compliance may also benefit from increased consumer confidence in the safety of their products.¹³⁶

The FDA has not yet released an estimate of the cost of compliance for food producers not qualifying for the exemption,¹³⁷ though presumably these numbers will accompany the release of FDA’s new rules on HACCP and produce standards.¹³⁸ Although the final FDA rules for preventive con-

2010 (2010), <http://www.cbo.gov/ftpdocs/120xx/doc12035/hr2751.pdf> (“[M]andatory recalls and risk-based preventive controls, could result in the assessment of civil or criminal penalties. Criminal fines are recorded as revenues, then deposited in the Crime Victims Fund, and later spent. Enacting H.R. 2751 could increase revenues and direct spending, but CBO estimates that the net budget impact would be negligible for each year.”).

¹³¹ *FDA Food Safety Modernization Act: Frequently Asked Questions*, FDA, <http://www.fda.gov/Food/FoodSafety/FSMA/ucm247559.htm> (last updated Nov. 10, 2011) (“The fees announced today allow FDA to recover 100% of its costs associated with certain domestic and foreign facility reinspections, failure to comply with a recall order, and certain importer reinspections. Previously, FDA bore the burden of these costs.”).

¹³² *Id.*

¹³³ Pollan & Schlosser, *supra* note 114.

¹³⁴ *FDA Food Safety Modernization Act: Frequently Asked Questions*, FDA, <http://www.fda.gov/Food/FoodSafety/FSMA/ucm247559.htm> (last updated Nov. 10, 2011).

¹³⁵ Pollan & Schlosser, *supra* note 114 (“[A] recent study by Georgetown University found that the annual cost of food-borne illness in the United States is about \$152 billion.”).

¹³⁶ See Michael Ollinger & Danna L. Moore, *The Direct and Indirect Costs of Food-Safety Regulation*, 31 REV. AGRIC. ECON. 247, 250 (2009) (While “[m]anagers make these investments to avoid costly recalls or other food-safety catastrophes, enhance their reputation with buyers, etc.,” compliance with FSMA could also enhance their reputation with buyers.).

¹³⁷ *FDA Food Safety Modernization Act: Frequently Asked Questions*, FDA, <http://www.fda.gov/Food/FoodSafety/FSMA/ucm247559.htm> (last updated Nov. 2, 2012) (“What are the estimated costs of a new inspection system – new inspectors, new processing, additional labs and reporting to Congress? What will the cost impact be on the farmer and consumer? It is too soon to know what the costs will be; FDA anticipates there will be some initial costs with the implementation of two rules that FDA anticipates releasing soon, the preventive controls and produce regulations.”).

¹³⁸ *Id.*

trols and produce safety standards are not yet complete,¹³⁹ the actual costs incurred by meat producers following implementation of the USDA HACCP rules in 2000 may provide a clue.¹⁴⁰ The USDA's ERS gathered studies highlighting the compliance costs of HACCPs relating to meat and poultry.¹⁴¹ The ERS cites a study by Michael Ollinger and Danna Moore, *The Direct and Indirect Costs of Food-Safety Regulation*, that analyzed "costs of sanitation and monitoring tasks, planning and reporting requirements, . . . testing mandates, and the relative costs of large and small plants."¹⁴² These regulations are very similar in design to those required by the FSMA, which mandates identification of foreseeable hazards, preventive controls at critical points, monitoring, and recordkeeping.¹⁴³

The Ollinger and Moore study attempts to reduce confusion about the actual costs of regulation on meat and poultry producers by using actual cost data collected from manufacturers themselves, the FSIS, and the Census.¹⁴⁴ The study analyzes costs of private actions—investments in food safety not required by regulation but made voluntarily by producers¹⁴⁵—as well as the direct and indirect costs of regulation.¹⁴⁶ Direct costs derive from the actual performance of HACCP tasks, such as planning and cleaning, while indirect costs derive from "the comparative advantage some plants have in meeting regulatory requirements (e.g., large plants may have lower per-unit regulatory costs because they can spread fixed regulatory costs over more volume)."¹⁴⁷

The results of Ollinger and Moore's study revealed that costs of private actions consistently exceeded direct costs of regulation.¹⁴⁸ It should be emphasized that these costs arose because "plants chose or were forced by their customers to go beyond" the FSIS standards.¹⁴⁹

As for the direct and indirect costs associated with the actual government regulations, the Ollinger and Moore study found that "economies of

¹³⁹ See *Food Safety Modernization Act: Implementation and Progress*, FDA, <http://www.fda.gov/Food/FoodSafety/FSMA/ucm250568.htm> (last updated Aug. 31, 2012).

¹⁴⁰ See Ollinger & Moore, *supra* note 136, at 247.

¹⁴¹ *Market Incentives & Government Regulation*, USDA, ECON. RESEARCH SERV., <http://www.ers.usda.gov/topics/food-safety/market-incentives-government-regulation/readings.aspx> (last updated May 26, 2012).

¹⁴² Ollinger & Moore, *supra* note 136, at 247.

¹⁴³ 21 U.S.C. § 350g (Supp. IV 2010).

¹⁴⁴ Ollinger & Moore, *supra* note 136, at 247-48.

¹⁴⁵ *Id.* at 250 ("Private actions include investments in human capital and innovative food-safety technologies and practices. . . . Other market-driven private actions include explicit agreements between plants and large buyers, such as fast food restaurant chains, in which plant managers agree to undertake food-safety process control tasks and make specific investments in return for guaranteed markets, higher volume orders, higher prices, or some other benefit.") (citation omitted).

¹⁴⁶ *Id.* at 248-50.

¹⁴⁷ *Id.* at 248.

¹⁴⁸ *Id.* at 260, 262.

¹⁴⁹ *Id.* at 249.

scale in food-safety process control give the very largest plants a substantial cost advantage over their smaller competitors.”¹⁵⁰ Despite this disparity in ability to absorb compliance costs, the study noted that existing small producers might minimize the effect by avoiding direct competition with larger producers.¹⁵¹ Small meat producers regulated by HACCP plans are at a disadvantage, but they have shown their businesses can survive by carving out their own niche markets.¹⁵²

Considering the current administration’s commitment to promoting growth in the small farm sector, however, the effect of HACCP-type regulations on market entry should also be considered.¹⁵³ Under the USDA regime for meat producers, existing producers possess a comparative advantage over new entrants because HACCP regulations raise the cost of entry for would-be producers.¹⁵⁴ If the same holds true under the new FSMA regulations, Secretary Vilsack’s goal of adding 100,000 new farmers to the economy would be pure fantasy in the absence of the Tester–Hagan exemptions for small farmers.¹⁵⁵

Ollinger and Moore also found the cost of complying with specific “process control tasks” mandated under HACCP regulations is significantly higher than simply setting performance standards and allowing food producers to meet the standards in the manner they think is best.¹⁵⁶ The new regulations under the FSMA include both HACCP-type tasks¹⁵⁷ and performance standards in the form of “science-based minimum standards for the safe production and harvesting of . . . fruits and vegetables.”¹⁵⁸

The principle of flexible compliance may seem less important under the FSMA than it is under the USDA rules because the Tester–Hagan Amendment creates essentially the same exemption for qualifying local food producers from both the § 350g preventive control requirements

¹⁵⁰ Ollinger & Moore, *supra* note 136, at 261.

¹⁵¹ *Id.* at 260 (“Large plants already enjoy substantial economies of scale, yet small plants persist by producing niche products and avoiding direct competition with their large competitors. Thus, the actual disproportionate impact on survival of the PR/HACCP on the survival of small plants relative to large ones may be quite small.”) (internal citation omitted).

¹⁵² *Id.* at 247, 260.

¹⁵³ Hamilton, *supra* note 3, at 127-28.

¹⁵⁴ Ollinger & Moore, *supra* note 136, at 251 (“[R]egulation favors incumbents because regulation raises industry entry costs.”).

¹⁵⁵ Hamilton, *supra* note 3, at 127-28.

¹⁵⁶ Ollinger & Moore, *supra* note 136, at 261-62 (“[T]he costs of complying with the generic *E. coli* and *Salmonella* performance standards was less than one-half the costs of performing SSOP and HACCP tasks in cattle and hog slaughter and raw and cooked meat. [This] finding means that if performance standards and process control tasks (SSOP and HACCP tasks) currently provide equal amounts of safety and if FSIS regulators wanted to enhance food-safety process control, then the same benefits at less than 40% the costs could be realized by raising the stringency of performance standards rather than the number of process control tasks.”).

¹⁵⁷ See generally 21 U.S.C. § 350g (Supp. IV 2010).

¹⁵⁸ 21 U.S.C. § 350h(a)(1) (Supp. IV 2010).

(process standards) and the § 350h standards for produce safety (performance standards).¹⁵⁹ However, a facility wishing to qualify for exemption from the § 350g HACCP-type process tasks must provide documentation demonstrating that it has identified potential health hazards and is adequately implementing its own controls, or it must provide documentation demonstrating that it is in compliance with state or local safety laws.¹⁶⁰ This essentially amounts to a flexible compliance standard for exemption from HACCP-type controls. As long as a qualifying local food producer provides the FDA with such documentation, it need not develop the precise preventive controls, monitoring systems, and recordkeeping systems mandated by the FSMA (all process standards).¹⁶¹

While the cost of providing this documentation is likely more than zero, it pales in comparison to what local food producers would face if they were not exempt from the new regulations, especially taking into account the disadvantage of their relative size.¹⁶² Therefore, in practice the Tester–Hagan exemptions to the FSMA process controls may act more as a flexible compliance option for small producers than as a total exemption. But allowing least cost compliance should still allow exempt producers to experience significant savings.¹⁶³

B. *Recommendations for Further Amendment to the FSMA*

Despite the potential compliance savings in direct and indirect costs to small farmers granted by the Tester–Hagan Amendment, the \$500,000 revenue constraint limits the Amendment’s economic protection of direct-to-consumer farmers. For example, Virginia’s Joel Salatin produces foods that have demonstrated greater inherent safety than industrial foods,¹⁶⁴ and he sells them only to local consumers—a practice the federal government clearly supports.¹⁶⁵ Yet Salatin—whose shell egg sales fall under the pur-

¹⁵⁹ Compare 21 U.S.C. § 350g(l) (Supp. IV 2010), with *id.* § 350h(f)(1) (both exemptions require more than half of sales to be to “qualified end-users” and cap producer revenue at \$500,000).

¹⁶⁰ 21 U.S.C. § 350g(l)(2) (Supp. IV 2010).

¹⁶¹ *Id.*

¹⁶² Taylor, *supra* note 15, at A531 (“The development, maintenance, and recordkeeping of HACCP plans is much more of a resource burden on small operators because of the economies of scale,” explains Mark Schad, a former small plant owner/operator who now works with other small operations to help them attain an FSIS grant of inspection. “There is not much difference in the cost associated with a HACCP plan whether an operator makes one hundred or one hundred thousand pounds of product.”).

¹⁶³ Unnevehr & Jensen, *supra* note 125, at 111 (“Among [Command and Control] approaches, process standards are less efficient than performance standards. . . . Setting performance standards and allowing choice of production methods, and over time, innovation to meet standards, should allow greater efficiency in meeting a particular public health goal.”).

¹⁶⁴ Taylor, *supra* note 15, at A531.

¹⁶⁵ MARTINEZ ET AL., *supra* note 1, at 35.

view of the FSMA and whose meat and poultry sales are subject to USDA-approved processing—would never qualify for exemption from HACCP-type controls even if the USDA employed the same exemption as the FDA because his operation grosses too much money.¹⁶⁶ The fact that the Tester–Hagan Amendment exemption does not extend to farmers like Salatin simply because of higher revenues conflicts with the spirit of the Amendment.¹⁶⁷ Moreover, setting a revenue ceiling for exemption from the FSMA’s costly compliance requirements creates a disincentive for local direct-to-consumer producers to expand production and may discourage new entrants in local farming.¹⁶⁸

While imposing a sales cap presents an obvious way to limit the reach of the Tester–Hagan exemptions,¹⁶⁹ it has the undesirable effect of undermining new entry and existing growth for farmers. This seems contrary to the policy objective, since the Senate’s stated purpose for including the Tester–Hagan Amendment was to protect local farming,¹⁷⁰ and the federal government has supported local food initiatives in several other ways.¹⁷¹ In recent years Congress has supported local food through legislation such as the Community Food Project Grants Program in the 1996 Farm Act,¹⁷² the

¹⁶⁶ See Gabor, *supra* note 47 (“Revenues this year will top \$2 million, nearly double the figure five years ago, says Salatin . . .”).

¹⁶⁷ DEMOCRATIC POLICY COMM., 111TH CONG., FOOD SAFETY AND AGRICULTURAL PRODUCERS 3 (Comm. Print 2010) (“The codification of produce safety standards would ensure that the future standards take into consideration sustainable agriculture and conservation practices; accommodate concerns about the scale of the operations; prevent impacts to organic agriculture; and provide flexibility to direct-to-consumer operations.”).

¹⁶⁸ See Hamilton, *supra* note 3, at 129-34 (“Singling out sales and primary occupation as the defining measure underpins an institutional bias against many farmers, new and old.”).

¹⁶⁹ 156 CONG. REC. S8266-67 (daily ed. Nov. 30, 2010) (statement of Sen. Thomas Harkin) (“I know that some of my colleagues think the Tester-sponsored language goes too far to help small growers and processors. I don’t think we have . . . There are some very important limitations on the Tester provisions in S. 510. First, small businesses as we define them here are really small—a company that does \$500,000 of sales a year is very small . . . The smallest member of the California League of Food Processors reports between \$2.5 and \$3 million a year in sales or five times as much as any company eligible under the Tester provisions.”).

¹⁷⁰ DEMOCRATIC POLICY COMM., 111TH CONG., FOOD SAFETY AND AGRICULTURAL PRODUCERS 3 (Comm. Print 2010) (“The codification of produce safety standards would ensure that the future standards take into consideration sustainable agriculture and conservation practices; accommodate concerns about the scale of the operations; prevent impacts to organic agriculture; and provide flexibility to direct-to-consumer operations.”); *see also* 156 CONG. REC. S8264-65 (daily ed. Nov. 30, 2010) (statement of Sen. Sheldon Whitehouse) (“I thank Senator Tester for his work on a compromise to protect farmers like those in Rhode Island, and throughout the Nation, who believe in the value of locally grown food.”).

¹⁷¹ MARTINEZ ET AL., *supra* note 1, at 35 (“Although the United States does not have a broad strategy of public procurement of local foods, there are policies and programs that support local food initiatives.”).

¹⁷² *Id.*

Child Nutrition and WIC Reauthorization Act of 2004,¹⁷³ and the 2008 Food, Conservation, and Energy Act.¹⁷⁴ Federal agencies have directly or indirectly supported local food, notably the Department of Defense¹⁷⁵ and the Centers for Disease Control and Prevention.¹⁷⁶ The USDA, under its own authority and marketing mandate, has created campaigns for local food such as the Community Food Security Initiative, the WIC Farmers' Market Nutrition Program, the Senior Farmers' Market Nutrition Program, the Federal State Marketing Improvement Program, the National Farmers' Market Promotion Program, the Specialty Crop Block Grant Program, and the Community Facilities Program.¹⁷⁷ In other words, 2009's Know Your Farmer Know Your Food campaign is the most direct USDA initiative supporting local food, but it is not the first.¹⁷⁸

Because the federal government has expressed clear intent that local agriculture should be promoted and supported, Congress should amend the FSMA by removing the gross sales limitation on the Tester–Hagan exemption. The “qualified end-user” requirement—that 50% or more of sales must be local¹⁷⁹—preserves the spirit of the Tester–Hagan Amendment, but the \$500,000 revenue limit discourages new entry into local farming and growth within that sector.¹⁸⁰ For existing producers at the margin, the costs of compliance associated with expanding their businesses such that revenues exceed \$500,000 may outweigh the benefits.¹⁸¹ The USDA's ERS reports that such uncertainty about the effect of regulatory regimes on business is a significant barrier for new entry into small-scale food production.¹⁸² Although the FDA has not yet announced the estimated costs of

¹⁷³ *Id.* (“requir[ing] school districts participating in federally funded meal programs to implement local wellness policies,” which “has led proponents to tout local foods as part of a healthy eating solution”).

¹⁷⁴ *Id.* at 38 (funding “the Business and Industry Guarantee Loan Program (B&I) to aid rural food enterprise entrepreneurs and local food distribution, and funding [] the Value-Added Agricultural Market Development (VAAMD) program emphasizing local food distribution”).

¹⁷⁵ *Id.* at 35 (“In 1994, the U.S. Department of Defense (DoD) began a project . . . referred to as the Fresh Program, partner[ing] with USDA to procure produce for institutions that was grown within their State, with preferences increasingly given to small and medium-sized farms.”).

¹⁷⁶ See *CDC's Healthy Communities Program*, CTR. FOR DISEASE CONTROL & PREVENTION, <http://www.cdc.gov/healthycommunitiesprogram/communities/> (last visited Jan. 2, 2013).

¹⁷⁷ See MARTINEZ ET AL., *supra* note 1, at 35-37.

¹⁷⁸ USDA, KNOW YOUR FARMER KNOW YOUR FOOD: OUR MISSION, http://www.usda.gov/wps/portal/usda/usdahome?navid=KYF_MISSION (last visited Jan. 2, 2013).

¹⁷⁹ 21 U.S.C. §§ 350g(1), 350h(f) (Supp. IV 2010).

¹⁸⁰ See Hamilton, *supra* note 3, at 129-34 (“Singling out sales and primary occupation as the defining measure underpins an institutional bias against many farmers, new and old.”).

¹⁸¹ Unnevehr & Jensen, *supra* note 125, at 123 (“[R]egulation has an impact on long-term incentives to invest in new technologies or inputs, and therefore is likely to bias the nature of productivity growth.”).

¹⁸² MARTINEZ ET AL., *supra* note 1, at iv, 27.

HACCP implementation in the industries subject to the FSMA,¹⁸³ HACCP regulations had significant direct and indirect costs when they were implemented in the meat and poultry industries, favoring the largest producers and processors.¹⁸⁴ Under the FSMA as written, when a potential new farmer considers market entry, or when an existing small farmer considers expansion, each must consider whether an operation with potential revenue of \$500,000 or more will provide benefits exceeding its costs, including HACCP.¹⁸⁵

If direct-to-consumer producers were not subject to the HACCP regulations of the FSMA, regardless of revenue, their food would not be any less safe; these food sources derive their safety largely from their traceability and transparency.¹⁸⁶ Direct-to-consumer producers would still have to provide documentation demonstrating the preventive measures they have adopted themselves,¹⁸⁷ and no producer is exempt from foodborne contaminant performance standards to be written by the FDA.¹⁸⁸ Moreover, if a producer initially qualifying for exemption from the HACCP regulations is identified as a source of food safety problems, the FDA has authority to withdraw its exemption.¹⁸⁹

By contrast, industrial producers will likely continue to find it too costly to market their products directly to consumers or to individual retail establishments, preferring instead to use wholesale distributors.¹⁹⁰ Thus, the “qualified end-user” requirement would likely still bar industrial producers

¹⁸³ FDA *Food Safety Modernization Act: Frequently Asked Questions*, FDA, <http://www.fda.gov/Food/FoodSafety/FSMA/ucm247559.htm> (last updated Nov. 10, 2011) (“It is too soon to know what the costs will be; FDA anticipates there will be some initial costs with the implementation of two rules that FDA anticipates releasing soon, the preventive controls and produce regulations.”).

¹⁸⁴ Ollinger & Moore, *supra* note 136, at 261 (“Results suggest that indirect and direct regulatory effects and private actions significantly affected food-safety costs. Some of the more notable findings are: (a) economies of scale in food-safety process control give the very largest plants a substantial cost advantage over their smaller competitors . . .”).

¹⁸⁵ See Richard Layard & Stephen Glaister, *Introduction to COST-BENEFIT ANALYSIS 1* (Richard Layard & Stephen Glaister eds., 2d ed., 1994) (“[I]t seems quite natural to refer to the ‘benefits of the next best alternative to A’ as the ‘costs of A.’ For if A is done those alternative benefits are lost. So the rule becomes: do A if its benefits exceed its costs, and not otherwise.”).

¹⁸⁶ Taylor, *supra* note 15, at A531 (Michael Pollan argues “that food safety problems from small players are ‘less catastrophic and easier to manage because local food is inherently more traceable and accountable.’”).

¹⁸⁷ 21 U.S.C. § 350g(l)(2) (Supp. IV 2010).

¹⁸⁸ *Id.* § 2201.

¹⁸⁹ *Id.* § 350g(l)(3) (“[I]f the Secretary determines that it is necessary to protect the public health . . . based on conduct or conditions associated with a qualified facility that are material to the safety of the food manufactured . . . the Secretary may withdraw the exemption provided to such facility . . .”).

¹⁹⁰ After all, it is the direct marketing to individual buyers that the USDA found to be one of the most significant costs to small farmers. MARTINEZ ET AL., *supra* note 1, at 30 (“[G]rowers who work off-farm generally have fewer incentives to expand and become more efficient than do small growers who do not participate in alternative, off-farm marketing activities.”).

from exemption.¹⁹¹ Because a distributor is not a restaurant and 21 U.S.C. § 350h(f)(4)(B) specifically says that a qualifying “consumer” does not include a business, the industrial producer would not find a loophole by removing the \$500,000 sales ceiling.¹⁹²

If the industrial producer is able to find a cost-effective means for selling at least 50% of its products to qualifying end-users, part of the safety rationale behind local food, particularly isolation and traceability, would be satisfied anyway. Imagine if a large industrial food conglomerate sold Salmonella-tainted spinach to 100 local restaurants. Because the conglomerate dealt directly with the restaurants instead of a network of distributors and wholesalers, the source of the Salmonella outbreak would be immediately identifiable. This is essentially the “built-in” safety advantage of local food. Moreover, the FDA retains the power to withdraw the HACCP exemption whenever it is necessary to protect the public health.¹⁹³

CONCLUSION

The FSMA attempts to deal with the tension between food safety and support for local agriculture, which often cannot bear the costs of increased safety measures such as HACCP.¹⁹⁴ The federal government recognizes the benefits of a strong local food presence by supporting marketing programs, such as Know Your Farmer Know Your Food,¹⁹⁵ yet Congress has mandated regulations designed for industrial food producers.¹⁹⁶ The Tester–Hagan Amendment to the FSMA largely corrects this problem by providing exemptions for certain local food producers from costly compliance. However, by maintaining a revenue ceiling as one of the requirements for exemption, the FSMA continues to distort incentives by hindering new farmers from entering the market and discouraging the expansion of existing direct-to-consumer farms. Congress can largely reconcile the economic interests of the local food movement with the stronger safety regulations of the FSMA by amending the Act to remove the Tester–Hagan Amendment’s revenue ceiling.

¹⁹¹ See 21 U.S.C. §§ 350g(l), 350h(f) (Supp. IV 2010).

¹⁹² *Id.* § 350h(f)(4)(B).

¹⁹³ *Id.* § 350g(l)(3).

¹⁹⁴ See 156 CONG. REC. S8266-67 (daily ed. Nov. 30, 2010) (statement of Sen. Thomas Harkin) (“[O]ne of the most difficult issues I have had to face as manager of S. 510 is the balance between small growers and processors and larger producers and food companies.”).

¹⁹⁵ USDA, News Release, *USDA Launches ‘Know Your Farmer, Know Your Food’ Initiative to Connect Consumers with Local Producers to Create New Economic Opportunities for Communities* (Sept. 15, 2009), <http://www.usda.gov/wps/portal/usda/usdahome?contentidonly=true&contentid=2009/09/0440.xml>.

¹⁹⁶ See POLLAN, *supra* note 2, at 249; Trexler, *supra* note 3, at 339-40; Breselor, *supra* note 17.









